

Evaluation of DSM-IV Personality Disorder Criteria as Assessed by the Structured Clinical Interview for DSM-IV Personality Disorders

Richard F. Farmer and Alexander L. Chapman

The primary objective of this research was to investigate the psychometric and diagnostic efficiency properties of DSM-IV personality disorder (PD) criteria as assessed by the Structured Clinical Interview for DSM-IV Personality Disorders (SCID-II). Logistic regression analyses were also employed to identify discriminating and nondiscriminating diagnostic criteria within specific PD categories. Results based on a community sample of 149 psychotropic medication-free persons, 58% of whom had at least one PD as determined from the SCID-II, suggest problems with the assessment and/or conceptualization of some PD cat-

egories, most notably obsessive-compulsive PD. For many PD concepts, diagnostic criteria were identified that either detracted from the overall internal consistency and diagnostic efficiency of their associated PD criteria set or failed to uniquely discriminate individuals with specific PDs from those without. Most of these findings cannot be clearly attributed to limitations associated with the method used to assess PD criteria.

Copyright 2002, Elsevier Science (USA). All rights reserved.

A RECENT REVIEW by Blashfield and Intoccia¹ of the personality disorder (PD) literature suggests that the rate of published articles on these conditions has slowed since the publication of the DSM-III.² Such a finding may seem counterintuitive, as formal recognition was first extended to these disorders in DSM-III through their inclusion on a separate dimension of clinical description (axis II), along with a listing of defining criteria and a specification of diagnostic decision rules. Although Blashfield and Intoccia do not speculate about possible reasons for the slowed post-DSM-III publication rate, one possible explanation may center on the growing dissatisfaction among researchers and clinicians with the DSM PD classification.³⁻⁶ While some of this dissatisfaction arises from theoretical and/or conceptual concerns, other sources include research findings that suggest quantitative problems with the DSM PD classification, most notably the validity of PD assessment.⁷

The relatively recent development of semistructured interviews has substantially enhanced the inter-rater and test-retest reliability of PD assessment⁸ and, consequently, has likely enhanced the validity of PD assessment. However, differences in content or question format across the various interviews used to assess PD concepts have likely resulted in poor diagnostic concordance, or the observation that an individual diagnosed as positive for a PD with one interview is often unlikely to be similarly diagnosed when an alternative interview measure is used.⁸⁻¹⁰ Related concerns have also been expressed about self-report inventories

used to assess and/or diagnose DSM PD concepts, as these measures have been reported to overdiagnose PDs and have very low diagnostic concordance with semistructured interviews.⁸⁻¹¹ Such findings pose a substantial challenge for the valid assessment of PDs, as well as the establishment of a clinical science of PDs.

Apart from the issue associated with the validity of PD assessment, a related concern is how PDs are defined. That is, the degree of validity associated with PDs is a joint function of the validity associated with the method used to assess PDs, as well as the validity of PD constructs themselves.¹² For example, poor discriminant validity among PD diagnostic criteria,^{13,14} and hence problems in the validity of PD constructs, may be an artifact of problems associated with the validity of the method used to assess these criteria. Unfortunately, given the current state of the science of PDs, it is presently most difficult if not impossible to completely disentangle the impact of the validity or utility of the assessment method from the validity or utility of the PD construct. Consequently, an examination of the psychometric aspects and util-

From the Department of Psychology, Idaho State University, Pocatello, ID.

Supported in part by Grant No. 849 from the Faculty Research Committee, Idaho State University.

Address reprint requests to Richard Farmer, Ph.D., Department of Psychology, University of Canterbury, Private Bag 4800, Christchurch, New Zealand.

Copyright 2002, Elsevier Science (USA). All rights reserved.
0010-440X/02/4304-0006\$35.00/0
doi:10.1053/comp.2002.33494

ity of DSM PD criteria must be undertaken with reference to characteristics of the measure used to assess these criteria, an issue not always considered when the psychometric features and utility of PD criteria have been examined.¹⁵

Although the present research literature does not clearly suggest an empirical advantage of one semistructured interview for assessing PDs over the existent possibilities, the widely used Structured Clinical Interview for DSM-IV Personality Disorders (SCID-II) and the associated SCID-II Personality Questionnaire (SCID-II-PQ)¹⁶ have been constructed so as to be conceptually and methodologically related and tightly linked to DSM criteria. In the typical use of the SCID-II assessment approach, the examinee first completes a self-report questionnaire (the SCID-II-PQ) by responding “yes” or “no” to a series of statements that correspond to each of the DSM-IV¹⁷ PD criteria (with the exception of a small number of criteria that are evaluated only in the course of the interview). Those items that are acknowledged as present on the SCID-II-PQ are then further explored during participation in a semistructured diagnostic interview (the SCID-II), where the examinee’s responses to interview questions are rated according to the following guidelines: ? = inadequate information, 1 = absent, 2 = subthreshold, and 3 = threshold. A rating of 3 is made only if the criterion in question meets or surpasses the DSM-IV–defined threshold for symptom presence and if the symptom is evaluated as pathological, persistent, and pervasive.¹⁸ Interview findings, in turn, are used in the determination of PD diagnoses, whereby the number of criterion features rated 3 for a specific PD must meet or exceed the threshold number of DSM criterion features that are necessary for diagnosis. Dimensional scores for each PD concept or PD grouping can also be derived from interview ratings.

One primary assumption of the two-stage approach to SCID-II assessment is that SCID-II-PQ items that are not endorsed (i.e., result in a “no” response) are likely to be similarly denied during the SCID-II. In support of this assumption, research on the diagnostic efficacy of the SCID-II approach has indicated that the false-negative rate associated with SCID-II-PQ PD diagnoses is very low (1.8% on average) across all PD concepts when referenced to SCID-II findings, although

there is a tendency for a moderately high rate of false-positives.^{19,20} The observation that items rated “no” on the SCID-II-PQ tend to be also rated as absent or subthreshold when examined during a diagnostic interview supports the validity of the first stage screening approach of SCID-II assessment.

Psychometrically, the SCID-II assessment procedure has both strengths and weaknesses, and appears to be comparable with other interview-based measures that purport to assess DSM PD concepts.⁷⁻⁹ Studies that have reported agreement in diagnostic decisions reached by two or more raters who used the SCID-II generally report good to very good inter-rater reliability.²¹⁻²⁵ The test-retest reliability of the SCID-II tends to be more moderate.^{26,27} For example, in one multisite study of 284 individuals,²⁶ kappa coefficients to index diagnostic stability were .53 for patient samples (range, .24 to .74) and .38 for nonpatient samples (range, .12 to .59), with the interval of time between separate interviews ranging from 1 day to 2 weeks. These findings are comparable to other studies that have examined the short-term stability of PD diagnoses derived from other semistructured interviews.²⁶ One published study examined the internal consistency of symptom ratings within PD concepts. This study,²⁵ which used an Italian translation of the SCID-II, reported very good internal consistency of PD symptoms within diagnostic categories.

In terms of the validity of the SCID-II, findings are somewhat equivocal. In a study of the SCID-II for DSM-III-R²⁸ PD concepts, Skodol et al.²⁹ found that diagnostic judgments derived from the SCID-II agreed moderately to highly with diagnostic decisions reached through use of Spitzer’s³⁰ LEAD standard (i.e., “longitudinal expert evaluation using all data”), with overall diagnostic power ranging between .45 to .95 across PD categories. However, fair to poor agreement has typically been observed when diagnostic decisions reached using the SCID-II were compared with those derived from other semistructured interviews.³¹⁻³⁵ For example, Skodol et al.³⁴ administered the SCID-II and the Personality Disorder Examination (PDE) to a sample of 100 psychiatric inpatients. The median kappa coefficient used to index the degree of diagnostic agreement across the 11 PDs assessed was .50 (range, .14 to .66). Similarly, SCID-II

diagnoses only share moderate to low associations with diagnostic decisions reached from self-report inventories that assess PDs.^{24,31,32,36} This latter finding is not unique to the SCID-II, as other semistructured interviews also tend to have poor rates of diagnostic concordance with self-report inventories.⁷

Research on the diagnostic efficiency of individual PD criterion features assessed with the SCID-II has been sparse. To our knowledge, only one published study, by Nurnberg et al.,³⁷ has reported such data. Nurnberg et al. also examined via stepwise logistic regression the ability of individual PD criteria to uniquely predict the presence versus absence of the diagnostic concept to which they belong. Such inquiries have potential utility for informing subsequent modifications in DSM PD criteria as they may suggest means to (1) increase the homogeneity of individuals diagnosed with particular PDs, and (2) eliminate criteria that do not discriminate between persons who do and do not meet diagnostic cut points. Nurnberg et al. concluded from their analyses that several DSM-III-R PD criteria could be eliminated in order to enhance diagnostic efficiency and increase the homogeneity of individuals within PD categories by decreasing the number of possible combinations of criteria necessary to meet maximally efficient diagnostic thresholds. However, in this study, measurement factors that may not be directly related to the utility of DSM PD criteria (i.e., inter-rater reliability associated with the diagnostic concept and its associated criteria) may better account for these conclusions. That is, poor performance of individual criteria within diagnostic concepts may have been the result unreliability of assessment. Diagnostic agreement among raters for many PD concepts in the Nurnberg et al. study tended to be moderate or modest (range of kappa for all 11 PDs, .82 to .35; median kappa = .61), and the inter-rater reliability of individual criteria was apparently not considered in the evaluation of criterion utility. Furthermore, five of the 11 PD concepts were infrequently observed in their sample (i.e., < 10 positive diagnoses); consequently, diagnostic efficiency statistics for criteria associated with these PD concepts may be somewhat unreliable due to the relatively low rate of PD occurrence.

With the exception of the Maffei et al.²⁵ study cited above that used an Italian translation, previ-

ous evaluations of SCID-II findings in the areas of internal consistency, diagnostic efficiency, and criterion utility have been performed on Version 1.0 of the SCID,³⁸ which corresponds to DSM-III-R PD criteria. Consequently, the primary objective of this research was to investigate the psychometric and diagnostic efficiency properties of DSM-IV PD criteria as assessed by the SCID-II. Interpretation of the properties of DSM-IV PD criteria will be made with reference to measurement characteristics associated with the method used to assess them (e.g., inter-rater reliability at the level of the individual criterion).

METHOD

Participants

Participants for this research were recruited for a study primarily designed to examine emotional and cognitive aspects of PD pathology. Persons were solicited from the greater southeastern Idaho community through newspaper advertisements, posted notices, and flyers, the content of which was similar to that used by Newman et al.³⁹ to recruit disinhibited subjects and by Farmer and Nelson-Gray⁴⁰ to recruit persons with PD pathology. Relevant portions of the three ads used in this research are as follows:

ARE YOU ADVENTUROUS OR IMPULSIVE? A Psychology faculty member at ISU is studying carefree persons who lead exciting, impulsive lives. If you are the type of person who is emotional, likes parties, craves excitement, takes chances, and tends to act on the spur of the moment, call (. . .).

ARE YOU ANXIOUS OR INHIBITED? A Psychology faculty member at ISU is studying persons who are anxious, shy, or apprehensive. If you are the type of person who avoids social activities, feels anxious or nervous, tends to trust their judgment rather than their feelings, and has difficulty making decisions, call (. . .).

ARE YOU ECCENTRIC OR UNIQUE? A Psychology faculty member at ISU is studying nonconforming persons who lead uncommon lives. If you are the type of person who is individualistic, self-reliant, socially apprehensive, original, or outlandish, call (. . .).

Potential subjects who responded to these ads were excluded during a phone prescreening in the event they were currently using psychotropic medication or if they were younger than 18 or older than 50 years of age. Prospective participants who did not meet these exclusionary criteria were provided with a detailed verbal description of the study specifics. Following verbal consent by potential participants, a packet of materials was sent to their homes that included an informed consent form. Only those participants who subsequently indicated an understanding of the contents of the consent form and who voluntarily signed it were included in this study.

All retained participants were paid either \$10 or \$20 based on their overall level of study participation. A total of 149 subjects (64 males, 85 females) define the sample employed in this

study. The mean age across participants was 28.21 years ($SD = 9.30$).

Assessment of PDs and traits. After consenting to participate in the research, subjects were mailed a variety of questionnaires to complete at their homes. Among these was the SCID-II-PQ,¹⁶ a 119 item self-report questionnaire that asked subjects to indicate the presence or absence (via a "yes" or "no" response) of PD features. Included within this set of questions were items that inquired about signs and symptoms associated with each of the 10 PDs coded on axis II of DSM (anxious-fearful cluster: avoidant [AVD], dependent [DEP], obsessive-compulsive [COM] PDs; odd-eccentric cluster: paranoid [PAR], schizotypal [SZT], and schizoid [SZD] PDs; erratic-dramatic cluster: histrionic [HST], narcissistic [NAR], borderline [BRD], antisocial [ANT] PDs), and two provisional DSM-IV PD concepts, passive-aggressive [PAG] and depressive [DPS] PDs. An additional seven items were added to the SCID-II-PQ to assess adult antisocial features relevant for the evaluation of the antisocial PD diagnosis. The content of these questions closely paralleled that contained in the SCID-II interview questions for these areas.

From initial SCID-II-PQ responses, the presence of PD features was subsequently assessed by the SCID-II¹⁶ by a licensed clinical psychologist (R.F.F.) with over 10 years experience of conducting research on PDs. Because previous research has indicated that the false-negative rate associated with nonendorsed SCID-II-PQ items relative to interview ratings is extremely low^{19,20}, and because the SCID-II PD assessment is typically conducted as a two-stage process, only those criterion features that were endorsed on the SCID-II-PQ by participants were further assessed during the SCID-II.

One third of all SCID-II interviews ($n = 50$) were reassessed by a second rater for purposes of establishing inter-rater reliability. Interviews used in the reliability assessments were randomly selected with the constraint that 75% ($n = 38$) of the interviews reviewed corresponded to individuals who met criteria for at least one PD. This was done so as to more thoroughly evaluate the reliability of PD assessment when PDs are present. The second rater, a doctoral candidate in clinical psychology (A.L.C.), was trained in the administration of the SCID-II by the first author. Ratings of subjects' responses to interview questions for reliability purposes were based on audiotaped recordings of the initial interview.

In the results presented below, the presence of PD categories was determined based on the SCID-II-PQ or SCID-II according to the categorical diagnostic decision rules outlined in DSM-IV. Dimensional scores based on SCID-II-PQ and SCID-II data for each of the PDs were also computed for each subject according to the following rules. For the SCID-II-PQ, items that were endorsed "yes" within PD categories were weighted with a value of 1.0 and were summed to yield an overall dimensional score ("no" responses were weighted 0.0). The resulting sum was then divided by the number of symptoms that defined the disorder, thus yielding a dimensional score. For the SCID-II, a score of 0.0 was assigned if a given criterion was evaluated as absent, 0.5 if the symptom was present but subthreshold (i.e., present but not to a clinically significant degree), and 1.0 if the symptom was present to a clinically significant degree. To create a dimensional interview score for each of the PDs, values assigned to symptom ratings for a single disorder were summed and then divided by the total number of symptoms for that

disorder. For example, if a given subject was determined to have three of the symptoms for dependent PD fully present, three partially present, and the remaining two absent, that person would receive a dimensional score of 0.56 for that disorder (i.e., $[3(1.0) + 3(0.5) + 2(0)]/8$). Cluster dimensional scores were computed in a similar manner, whereby dimensional scores that corresponded to PD concepts within a cluster were summed and then divided by the number of PD concepts within the cluster. To arrive at an PD dimensional score to index the degree of overall PD pathology, dimensional scores for all PD concepts (including the provisional PDs) were summed and then divided by the total number of PDs assessed (i.e., 12).

RESULTS

Inter-rater Reliability of PD Diagnoses, Symptom Ratings, and Dimensional Scores

Table 1 presents the reliability estimates for each PD concept and PD groupings when the data are categorical (i.e., diagnostic judgments, individual symptom ratings) and dimensional (i.e., based on scores that indicate the degree of PD pathology present). Kappa coefficients (κ) are only presented for individual PD diagnostic concepts that were diagnosed a total of 10 or more times across raters in order to avoid the potential inflation, deflation, and/or unreliability of the kappa statistic that can occur when few positive cases are observed. That is, the sum of all positive diagnoses for given PD diagnostic concept was derived for both raters, and then divided by 2. A resulting value ≥ 5 indicates that a positive diagnosis for that concept was made at least 10 times across both raters.

Reliability of summary dimensional scores for PD concepts or groupings across subjects was evaluated via intraclass correlation coefficients (ICC),⁴¹ which are conceptually similar to weighted kappas in the measurement of reliability.⁴² In the interpretation of kappa and ICC coefficients, values of 0.75 or greater are indicative of excellent agreement beyond chance, those less than 0.75 but greater than 0.40 are reflective of good to fair agreement beyond chance, and those falling below .40 are regarded as indicative of poor agreement.⁴³

Overall, inter-rater reliability was excellent. For diagnostic judgments based on the 10 PDs delineated on axis II of DSM-IV that occurred with sufficient frequency ($n = 6$ of 10), the median kappa value was .902 (range, .852 for NAR to .947 for AVD). Kappa values for the provisional PD concepts of PAG and DPS PDs were .912 and .940,

Table 1. Reliability of Personality Disorder Diagnoses, Symptom Ratings, and Dimensional Scores

Diagnostic Category	N*	Kappa		Intraclass Correlations
		Category (presence/absence)	Ratings of Individual Symptoms†	Dimensional Scores
Anxious-fearful cluster	24	.920	.921	.993
Avoidant	12.5	.947	.931 (.859-1.00)	.994
Dependent	1	—	.912 (.695-1.00)	.986
Obsessive-compulsive	16	.908	.900 (.754-1.00)	.973
Erratic-dramatic cluster	17.5	.869	.804	.980
Histrionic	1.5	—	.775 (.306-.881)	.918
Narcissistic	8	.852	.664 (.582-.798)	.966
Borderline	13	.896	.860 (.496-.968)	.971
Antisocial	4	—	.846 (.458-1.00)	.979
Odd-eccentric cluster	15.5	.860	.845	.881
Paranoid	11	.883	.900 (.820-.968)	.987
Schizotypal	6.5	.912	.720 (.110-.893)	.873
Schizoid	2	—	.583 (.489-.921)	.912
Provisional PD concepts				
Passive-aggressive	6.5	.912	.939 (.881-.962)	.985
Depressive	10.5	.940	.963 (.922-1.00)	.996
Any PD	37	.896	—	—
All PDs	—	—	.869	.934

*N is the total number of times that a given diagnosis was made by both interviewers divided by 2. Kappa coefficients are only presented for those diagnostic categories where this value is ≥ 5 .

†Values in this column correspond to the median kappa coefficient for ratings of individual symptoms within each PD category. Value ranges within parentheses indicate lowest and highest kappa values associated with specific symptoms within PD criterion sets.

respectively. At the cluster level of analysis, corresponding kappa values were .920 for the anxious-fearful cluster, .869 for the erratic-dramatic cluster, and .860 for the odd-eccentric cluster. Finally, kappa for diagnostic agreement on the presence versus absence of any PD (including provisional PDs) across subjects was .896. For ratings of individual criteria across all PD concepts, the overall median kappa was .869 (range, .583 for SZD to .963 for DPS).

For dimensional summary scores, the median ICC value for the 10 axis II PDs was .972 (range, .873 for SZT to .994 for AVD). ICCs for the provisional PD concepts of PAG and DPS PDs were .985 and .996, respectively. For the PD clusters, ICCs were .993 for the anxious-fearful cluster, .980 for the erratic-dramatic cluster, and .881 for the odd-eccentric cluster. Finally, the ICC that corresponded to the degree of agreement in dimensional score ratings based on all PD features (including those for provisional PDs) was .934.

When the reliability of PD assessment was examined as a function of subject sex, few differences were noted. In terms of diagnostic reliability

for specific PD concepts, NAR displayed the greatest disparity, where diagnostic agreement was greater for men ($n = 23$; $\kappa = 1.00$) compared to women ($n = 27$; $\kappa = .630$). In the vast majority of cases, the reliability of individual categorical symptom ratings was comparable among male and female participants. Among instances where κ could be computed for individual symptoms for both sexes (81 of 108 symptoms, or 75%), only four PD symptoms had κ values that differed by a value of .30 or more. These were PAR criterion 7 ("doubts fidelity", $\kappa = .639$ for males, 1.00 for females); SZD criterion 7 (emotional coldness; $\kappa = .740$ for males, $-.059$ for females), HST criterion 7 (suggestibility; $\kappa = 1.00$ for males, .690 for females), and BRD criterion 3 (unstable self-image; $\kappa = .596$ for males, .933 for females). Separate ICCs as a function of subject sex that corresponded to dimensional PD scores also suggested relatively modest sex differences in reliability. The greatest disparity in ICC coefficients among males and females for specific diagnostic concepts occurred for SZT (.917 for females, .941 for males).

Table 2. DSM-IV Personality Disorder Diagnoses and Dimensional Score Associations Based on SCID-II-PQ and SCID-II Data

Diagnostic Concept	PD Diagnosis					
	Frequency (and rate) of Diagnosis		Agreement Indices			
	SCID-II-PQ	SCID-II	κ	FP	PPP	r^*
Anxious-fearful cluster	108 (72.5)	46 (30.9)	.29	.60	.43	.88
Avoidant	54 (36.2)	21 (14.1)	.45	.26	.39	.88
Dependent	9 (6.0)	1 (0.7)	.19	.05	.11	.84
Obsessive-compulsive	90 (55.7)	33 (22.1)	.31	.49	.37	.79
Erratic-dramatic cluster	44 (29.5)	44 (29.5)	1.00	.00	1.00	.92
Histrionic	16 (10.7)	8 (5.4)	.64	.06	.50	.84
Narcissistic	70 (47.0)	19 (12.8)	.28	.39	.27	.84
Borderline	68 (45.6)	24 (16.1)	.37	.35	.35	.90
Antisocial	32 (21.5)	15 (10.1)	.58	.13	.47	.94
Odd-eccentric cluster	87 (58.4)	30 (20.1)	.30	.48	.35	.87
Paranoid	58 (38.9)	19 (12.8)	.37	.30	.33	.93
Schizotypal	44 (29.5)	14 (9.4)	.28	.24	.28	.74
Schizoid	38 (25.5)	5 (3.4)	.18	.23	.13	.76
Provisional PD concepts						
Passive-aggressive	35 (23.5)	12 (8.1)	.44	.17	.34	.91
Depressive	36 (24.2)	20 (13.4)	.66	.12	.56	.97
Any PD	128 (85.9)	86 (57.7)	.37	.67	.67	—
All PDs	—	—	—	—	—	.98

Abbreviations: SCID-II, Structured Clinical Interview for DSM-IV Personality Disorders; SCID-II-PQ, SCID Personality Questionnaire; κ , kappa coefficient of diagnostic agreement; FP, false-positive rate of SCID-II-PQ findings relative to SCID-II findings; PPP, positive predictive power values for the SCID-II-PQ relative to the SCID-II. r , Pearson product-moment correlation based on SCID-II and SCID-II-PQ dimensional scores.

*All r correlations are significant at $P < .001$.

Frequency of DSM-IV PD Diagnoses and Dimensional Rating Covariations as a Function of SCID-II-PQ and SCID-II Assessment

Table 2 presents SCID-II-PQ and SCID-II data for (1) the rates of PD diagnoses based on DSM-IV diagnostic decision rules; (2) unweighted kappa statistics to index the degree of diagnostic concordance shared by these two assessment methods; (3) the false-positive rate associated with SCID-II-PQ assessment relative to SCID-II findings; (4) positive predictive power (PPP) values to index the degree of diagnostic efficiency of the SCID-II-PQ relative to the SCID-II; and (5) Pearson correlations (r) to assess the degree to which SCID-II-PQ and SCID-II dimensional scores share linear associations.

As noted in Table 2, all PD concepts were well represented (i.e., diagnosed) when DSM-IV diagnostic decision rules were applied to the SCID-II-PQ. A notable exception is DEP PD, which was only diagnosed among 6% of the sample. Similarly, DEP was infrequently observed (0.7% of sample) when SCID-II findings were used to determine diagnoses.

Table 2 indicates that the false-positive rate for diagnosis associated with the SCID-II-PQ relative to SCID-II findings is quite high, an expected finding given the screening function of the SCID-II-PQ in the two-stage assessment process. When DSM-IV PD diagnostic decision rules were applied to SCID-II-PQ responses, the rate for any PD diagnosis in the sample (85.9%) substantially exceeded that based on SCID-II findings (57.7%), resulting in an overall false-positive rate of 67%. Indeed, for many of the specific PD concepts, the false-positive rate (i.e., a ratio derived by dividing the number of times the SCID-II-PQ diagnosed a PD as positive when it was actually negative as referenced to SCID-II findings by the total number of negative SCID-II diagnoses for that PD concept) is fairly high (e.g., COM, NAR, BRD, PAR); however, false-positives are infrequently observed among other PD concepts (e.g., DEP, HST, ANT, PAG, and DPS). At the cluster level of analysis, a striking finding is that the false-positive rate for any erratic-dramatic personality disorder (i.e., HST, NAR, BRD, and ANT) is 0%. Kappa coefficients for specific PD concepts suggest that, with

the possible exceptions of HST, DPS, and ANT, diagnostic decisions reached by SCID-II-PQ assessment have little association with those obtained from SCID-II assessment. Similarly, SCID-II-PQ derived diagnostic judgments at the cluster level of analysis have little association to SCID-II findings with one notable exception: for diagnoses within the erratic-dramatic cluster. As indexed by kappa, diagnostic decisions derived from SCID-II-PQ and SCID-II assessments are very highly related at the cluster level of analysis for erratic-dramatic PD concepts. Some caution must be exercised, however, in the interpretation of kappa values that correspond to SCID-II-PQ and SCID-II diagnostic concordance as these values are likely inflated to a small degree given that the study design excluded the false-negative cell (i.e., SCID-II-PQ negatives but SCID-II positives).

In order to determine degree of diagnostic concordance between the SCID-II-PQ and SCID-II, PPP values were also examined. PPP presented here represents the value derived from of the number of true-positive cases for a diagnosis (i.e., when the diagnosis was made based on the SCID-II) where the SCID-II-PQ also indicated the positive presence of a diagnosis divided by all positive results derived from the SCID-II-PQ. As Table 2 indicates, positive diagnoses derived from the SCID-II-PQ did not predict well the presence of specific SCID-II diagnoses; however, the SCID-II-PQ did a reasonably good job predicting the presence of any PD diagnosis (67%) and an exceptional job predicting the positive presence of any erratic-dramatic PD (100%). That is, for this sample, there was perfect agreement between the SCID-II-PQ and SCID-II for the positive presence of any erratic-dramatic cluster diagnosis, whereby if an individual met diagnostic thresholds for any erratic-dramatic PD on the SCID-II-PQ, he or she also met diagnostic criteria for at least one erratic-dramatic PD based on the SCID-II. When this finding is considered in conjunction with associated kappa and false-positive rates, an implication is that the SCID-II-PQ may be an excellent screening tool for the presence of any erratic-dramatic PD concept.

Table 2 also reports Pearson correlations that index the degree of association between dimensional scores SCID-II and SCID-PQ for PD diagnostic concepts and categories. These correlations

reveal strong linear associations ($r_s \geq .80$) between SCID-II and SCID-II-PQ dimensional scores, with the exception of COM ($r = .79$), SZD ($r = .76$), and SZT ($r = .76$). The absence of a strong linear association among the set of scores for SZT might have been anticipated, however, as three of the nine criteria that define this PD describe behavior observed during the SCID-II and, as such, are not assessed at all by the SCID-II-PQ.

Patterns of PD Comorbidity and Covariation Based on SCID-II Data

Based on SCID-II findings, a total of 86 individuals (57.7% of sample) met criteria for at least one PD. Persons diagnosed with a PD from the SCID-II had an average of 2.22 PDs (i.e., 191 PDs diagnosed overall for 86 individuals who were diagnosed with at least one PD; range of comorbid PD frequency among positive cases was one to eight PDs). Only one significant sex difference was noted in the percentage of cases diagnosed with specific PDs, and this occurred for ANT, $\chi^2(1) = 6.28, P < .05$. Males were significantly more likely to be diagnosed with ANT than females (17.2% and 4.7% of sample, respectively). When mean dimensional scores corresponding to specific PD concepts were considered as a function of sex, more significant differences were noted. These occurred for SZT, $t(144) = 2.13, P < .05$; NAR, $t(147) = 2.60, P < .05$; and ANT, $t(101.37) = 4.15, P < .001$. In each case, mean dimensional scores were higher for males than females (mean = .34 and .27 for SZT; .34 and .24 for NAR; and .20 and .11 for ANT).

To illustrate patterns of diagnostic comorbidity in PD diagnosis derived from SCID-II findings, the upper half of Table 3 presents the percentage of persons who met criteria for each PD given a positive diagnosis for another PD. As Table 3 illustrates, PAG and BRD displayed generally moderate to high levels of comorbidity (> 40%) with other PD concepts, a finding that is consistent with previous research.⁴⁴ Whereas BRD is often observed when another disorder is present (particularly HST, ANT, SZT, and PAG), the co-occurrence of these and other PDs tends to be relatively infrequent when BRD is present. Conversely, the PAG diagnosis is relatively infrequent when other PDs are present; however, other PDs commonly co-occur (particularly COM, NAR, BRD, and PAR) when PAG is present. Also noteworthy is the co-

Table 3. Patterns of Personality Disorder Diagnostic Comorbidity and Symptom Covariation

Diagnostic Category	Diagnostic Comorbidity Probabilities*											
	AVD	DEP [†]	COM	HST	NAR	BRD	ANT	PAR	SZT	SZD	PAG	DPS
Avoidant (AVD)	—	—	.33	.00	.00	.24	.00	.24	.19	.10	.10	.43
Dependent (DEP) [†]	—	—	—	—	—	—	—	—	—	—	—	—
Obsessive-compulsive (COM)	.15	—	—	.03	.18	.24	.15	.18	.12	.03	.18	.30
Histrionic (HST)	.00	—	.13	—	.50	.50	.38	.25	.25	.00	.38	.00
Narcissistic (NAR)	.00	—	.32	.21	—	.42	.32	.26	.21	.00	.32	.05
Borderline (BRD)	.21	—	.33	.17	.33	—	.33	.33	.29	.08	.29	.29
Antisocial (ANT)	.00	—	.33	.20	.40	.53	—	.27	.27	.00	.33	.13
Paranoid (PAR)	.26	—	.32	.10	.26	.42	.21	—	.21	.05	.42	.37
Schizotypal (SZT)	.29	—	.29	.14	.27	.50	.27	.27	—	.27	.21	.29
Schizoid (SZD)	.40	—	.20	.00	.00	.40	.00	.20	.80	—	.20	.60
Passive-aggressive (PAG)	.17	—	.50	.25	.50	.58	.42	.67	.25	.08	—	.42
Depressive (DPS)	.43	—	.48	.00	.05	.35	.10	.35	.20	.15	.24	—

	Pearson Product-Moment Correlations [‡]											
	AVD	DEP	COM	HST	NAR	BRD	ANT	PAR	SZT	SZD	PAG	DPS
Avoidant (AVD)	—	.34	.34	-.26	-.06	.31	-.07	.42	.34	.40	.32	.63
Dependent (DEP)		—	.27	.25	.35	.29	.06	.41	.30	.11	.36	.33
Obsessive-compulsive (COM)			—	.00	.12	.23	.02	.25	.19	.21	.32	.24
Histrionic (HST)				—	.54	.29	.26	.12	.20	-.19	.20	-.04
Narcissistic (NAR)					—	.38	.37	.38	.32	.08	.52	.17
Borderline (BRD)						—	.48	.57	.49	.25	.66	.62
Antisocial (ANT)							—	.34	.34	.10	.38	.18
Paranoid (PAR)								—	.52	.30	.57	.61
Schizotypal (SZT)									—	.47	.47	.40
Schizoid (SZD)										—	.41	.47
Passive-aggressive (PAG)											—	.57
Depressive (DPS)												—

*Values above and below the diagonal indicate probability of a column diagnosis given the presence of a row diagnosis.

[†]Data are not presented for dependent PD as this diagnosis was made only once.

[‡]Correlations $\geq .23$ or $\leq -.23$ are significant at $P \leq .01$. Correlations $.23 > r \geq .17$ or $-.17 \leq r < -.23$ are significant at $P \leq .05$.

morbidity of SZT (80%) and DPS (60%) among persons with SZD, the comorbidity of BRD (53%) among persons with ANT, and comorbidity that NAR (50%) and BRD (50%) have among persons with HST. All other comorbidity percentages are less than 50%.

Patterns of PD dimensional score covariation were also examined. The lower half of Table 3 presents separate sets of Pearson product-moment correlations for PD dimensional scores based on SCID-II ratings. When correlational data are considered, the magnitude of dimensional score covariation across PD concepts are greatest for PAG, PAR, BRD, and DPS.

Internal Consistency of PD Criteria

This and the following two sections contain descriptions of psychometric and diagnostic efficiency analyses for individual PD concepts. Beginning with the psychometric evaluation of PD

criteria, Table 4 presents internal consistency estimates based on interviewer ratings for specific PD concepts as assessed by the SCID-II. As indicated in Table 4, internal consistency for criteria within specific PD concepts is $\geq .60$ for all PD concepts, with the exception of criteria that correspond to COM ($\alpha = .53$). Furthermore, internal consistency is enhanced only to a very small degree when detractor criteria are removed.

When the internal consistency of PD criterion sets was considered as a function of subject sex, similar patterns for both groups tended to be the rule. The greatest discrepancies occurred for HST ($\alpha = .56$ for females, $.68$ for males) and the conduct disorder items of ANT ($\alpha = .60$ for females, $.78$ for males). At the level of individual symptoms, some disparities in symptom performance relative to the overall diagnostic concept were noted as a function of sex. DEP criterion 6 (helpless when alone) was a negative discriminator (i.e.,

Table 4. Internal Consistency Analyses of Personality Disorder Criterion Sets

Diagnostic Concept	No. of Criteria*	α	Detractors†
Avoidant	7	.86	(none)
Dependent	8	.61	5,7 (.66)
Obsessive-compulsive	8	.53	5,7 (.56)
Histrionic	8	.65	5 (.68)
Narcissistic	9	.76	(none)
Borderline	9	.79	(none)
Antisocial			
Conduct disorder	15	.75	15 (.76)
Adult antisocial	7‡	.72	6,7 (.74)
Paranoid	7	.75	(none)
Schizotypal	9	.69	6,4 (.71)
Schizoid	7	.63	(none)
Passive-aggressive	7	.67	(none)
Depressive	7	.81	5 (.84)

*Only the highest rated item among multiple items for a single criterion was used in the computation of dimensional scores and in the consideration of PD diagnostic judgments.

†Detractors = criterion number of DSM-IV PD concept that refers to corresponding SCID-II item that detracts from overall internal consistency of PD concept. Detractors were sequentially removed based on the magnitude of the corrected item-to-total correlation, with the item with the lowest correlation removed first followed by a recomputation of the reliability coefficient and item-to-total correlations. Values in parentheses under this column indicate adjusted internal consistency estimate once all detractor items had been removed.

‡An additional seven items were added to the questionnaire to assess adult antisocial features relevant for the evaluation of the antisocial PD diagnosis. The content of these questions closely paralleled that contained in the SCID-II-based questions for these areas.

had a negative corrected item-to-total correlation with the dimensional score for this concept) for males ($r = -.02$), but not for females ($r = .46$). Similarly, ANT conduct disorder criterion 4 (physically cruel to people) was a negative discriminator for females ($r = -.03$), but not males ($r = .43$).

Sensitivity, Specificity, and Predictive Power of Individual PD Criterion Features in Relation to Their Associated PD Category as Based on SCID-II Ratings

For analyses presented in this and the following section, SCID-II ratings were transformed so as to represent a dichotomous assessment of criterion features. These transformations were performed as these weightings are consistent with the practice of DSM PD diagnostic determination. Criteria that were previously weighted 0.0 (for absent) or 0.5 (for present but subthreshold) were weighted 0.0 to

indicate the failure to reach diagnostic thresholds that indicated criterion presence. Items that were previously weighted 1.0 (present to a clinically significant degree) retained this weighting for the following analyses. Data for PD concepts with unknown diagnostic reliability in relation to this study and that occurred with low frequency ($\leq 10\%$) are not presented (i.e., DEP, SZD, HST, ANT).

All of the analyses presented in this and the following sections are based on data from both males and females. Separate analyses for the evaluation of diagnostic efficiency as a function of subject sex are not advisable for this sample for two reasons: (1) the resultant lower rate of positive diagnoses for specific PD categories when examined separately by sex and, consequently, less reliable findings; and (2) substantially reduced statistical power to unacceptable levels when criteria within PD concepts are used as predictors of positive and negative PD diagnoses.

Table 5 presents sensitivity, specificity, PPP, and negative predictive power values for individual PD criteria in relation to their associated PD category. Sensitivity (SEN) as used here refers the probability that the criterion is evaluated as meeting or exceeding clinical thresholds (weighted 1.0) given a positive diagnosis for its associated PD category. Specificity (SPC) refers to the probability that a criterion is evaluated as occurring below clinical thresholds (weighted 0.0) given the absence of a diagnosis for its associated PD category. PPP refers to the ratio to true-positive cases where the criterion is rated as present (i.e., instances where a criterion is rated as clinically significant *and* the diagnosis is positive) to all cases when the criterion is rated as positive (i.e., all cases for which the criterion is rated as meeting or exceeding clinical thresholds). Finally, negative predictive power (NPP) refers to the ratio of true-negative cases where the criterion is rated as absent (i.e., instances where a criterion is rated as subthreshold or absent *and* there is no diagnosis) to all cases when the criterion is rated as negative (i.e., all instances where the criterion is rated as subthreshold or absent). Additional descriptions of these concepts are provided by Baldessarini et al.⁴⁵

Across PD concepts, PPP values generally tended to be the lowest ($\leq .50$) among the four diagnostic efficiency values examined, whereas

Table 5. Sensitivity, Specificity, Positive Predictive Power, and Negative Predictive Power Indices for and Logistic Regression of PD Criteria Assessed by the SCID-II

Diagnostic Concept (and criterion label)	Criterion No.	Diagnostic Efficiency Statistics				Step Entered	Logistic Regression Analyses		
		SEN	SPC	PPP	NPP		χ^2	Significance (<i>P</i>)	Classification
Avoidant (AVD)									
(avoids activities)	AVD1	.67	.95	.70	.95	4	7.95	.005	99.3
(noninvolvement)	AVD2	.67	.95	.67	.95				
(restraint)	AVD3	.48	.88	.40	.91	3	19.34	<.001	98.6
(fear of criticism)	AVD4	.91	.83	.46	.98				
(socially inhibited)	AVD5	.86	.94	.72	.98	2	21.34	<.001	95.9
(low self view)	AVD6	.86	.91	.60	.98				
(non-risk-taking)	AVD7	.81	.97	.81	.97	1	68.13	<.001	95.2
	Means	.75	.92	.62	.96				
Obsessive-compulsive (COM)									
(details/order)	COM1	.46	.87	.50	.85	5	11.71	.001	94.6
(perfectionism)	COM2	.61	.91	.64	.89	1	35.38	<.001	83.9
(work devotion)	COM3	.48	.90	.57	.86	4	22.66	<.001	91.3
(morality/ethics)	COM4	.46	.88	.52	.85	7	17.58	<.001	99.3
(retains objects)	COM5	.58	.81	.46	.87	6	17.13	<.001	96.0
(nondelegation)	COM6	.73	.81	.52	.91	2	25.51	<.001	84.6
(miserly)	COM7	.12	.95	.40	.79	8	5.00	.025	100.0
(rigidity/stuborn)	COM8	.97	.53	.37	.98	3	22.59	<.001	87.9
	Means	.55	.83	.50	.88				
Paranoid (PAR)									
(suspicion)	PAR1	.89	.81	.40	.98	4	12.70	<.001	95.8
(doubts loyalty)	PAR2	.68	.91	.52	.95	2	15.01	<.001	91.6
(nonconfiding)	PAR3	.89	.82	.42	.98	1	36.71	<.001	92.3
(reads meaning)	PAR4	.47	.83	.29	.91				
(bears grudges)	PAR5	.68	.88	.46	.95	3	13.50	<.001	92.3
(perceives attacks)	PAR6	.42	.91	.40	.92	5	17.02	<.001	97.2
(doubts fidelity)	PAR7	.44	.92	.44	.92	6	10.54	.001	99.3
	Means	.64	.87	.42	.94				
Schizotypal (SZT)									
(ideas of reference)	SZT1	.85	.73	.23	.98	2	15.50	<.001	93.2
(odd beliefs)	SZT2	.77	.82	.29	.97	4	12.96	<.001	96.6
(odd perceptions)	SZT3	.85	.82	.31	.98	1	27.23	<.001	90.4
(odd thinking)	SZT4	.54	.84	.25	.95				
(suspiciousness)	SZT5	1.00	.53	.17	1.00				
(odd affect)	SZT6	.46	.89	.29	.94	5	10.94	.001	97.9
(odd behavior)	SZT7	.46	.94	.43	.95				
(no friends)	SZT8	.54	.92	.39	.95	3	15.22	<.001	93.8
(social anxiety)	SZT9	.31	.97	.50	.94				
	Means	.64	.83	.32	.96				
Narcissistic (NAR)									
(grandiosity)	NAR1	.53	.92	.50	.93	3	14.04	<.001	95.3
(fantasy proneness)	NAR2	.74	.82	.37	.96	4	11.46	.001	96.6
(self-view: special)	NAR3	.74	.94	.64	.96	1	42.74	<.001	91.3
(needs admiration)	NAR4	.58	.85	.37	.93				
(entitlement)	NAR5	.79	.92	.58	.97	2	19.89	<.001	93.3
(exploitative)	NAR6	.79	.91	.56	.97	5	8.44	.004	97.3
(lacks empathy)	NAR7	.74	.87	.45	.96				
(envious)	NAR8	.84	.65	.26	.97				
(arrogant)	NAR9	.68	.85	.39	.95	6	13.33	<.001	99.3
	Means	.71	.86	.46	.96				
Borderline (BRD)									
(avoid abandonment)	BRD1	.38	.93	.50	.89	3	17.43	<.001	92.5
(unstable relations)	BRD2	.67	.81	.40	.93				
(unstable self-image)	BRD3	.79	.77	.40	.95	4	10.52	.001	95.9
(impulsivity)	BRD4	.79	.72	.35	.95	5	11.10	.001	96.6
(suicidal behavior)	BRD5	.33	.94	.53	.88				
(affective instability)	BRD6	.92	.83	.51	.98	1	50.12	<.001	83.7
(emptiness)	BRD7	.75	.85	.49	.95	2	17.83	<.001	94.6
(anger)	BRD8	.79	.84	.49	.96				
(paranoia/dissociate)	BRD9	.58	.82	.39	.91	6	16.57	<.001	98.6
	Means	.67	.83	.45	.93				
Passive-aggressive (PAG)									
(task resistance)	PAG1	.67	.79	.21	.96				
(misunderstood)	PAG2	.83	.72	.20	.98				
(argumentative)	PAG3	.58	.88	.29	.96	4	7.62	.006	98.0
(scorns authority)	PAG4	.67	.87	.31	.97	3	13.85	<.001	98.0
(envy/resentment)	PAG5	.83	.91	.46	.98	2	20.57	<.001	96.0
(misfortune)	PAG6	.67	.89	.35	.97	5	9.44	.002	99.3
(defiance/contrition)	PAG7	.58	.98	.70	.96	1	28.18	<.001	94.6
	Means	.69	.86	.36	.97				
Depressive (DPS)									
(low mood)	DPS1	.80	.91	.57	.97	2	18.79	<.001	95.3
(low self-esteem)	DPS2	.95	.89	.58	.99	1	61.09	<.001	89.9
(self-critical)	DPS3	.95	.75	.37	.99				
(worry)	DPS4	.95	.75	.37	.99	3	16.10	<.001	96.0
(other critical)	DPS5	.45	.74	.21	.90				
(pessimism)	DPS6	.65	.92	.54	.94	4	7.81	.005	98.0
(guilt/remorse)	DPS7	.95	.83	.46	.99				
	Means	.81	.83	.44	.97				

Abbreviations: SCID-II, Structured Clinical Interview for DSM-IV Personality Disorders; Criterion No., corresponding PD criterion number as specified in DSM-IV; SEN, sensitivity; SPC, specificity; PPP, positive predictive power; NPP, negative predictive power; Classification, percentage of cases accurately classified as a result of the stepwise logistic regression model.

SPC and NPP values were generally the highest ($\geq .80$), with NPP values tending to be slightly higher than SPC values. This general pattern of findings is similar to those reported by Nurnberg et al.³⁷ for DSM-III-R PD criteria. Overall, Table 5 indicates (1) that criterion features are often rated as absent when the diagnostic concept to which the criterion features belong is also rated as absent (SPC); (2) that criteria are typically rated as absent among true-negative cases for a given diagnosis, and not rated as negative when the diagnosis is present (NPP); (3) that sensitivity (SEN) values (i.e., proportion of instances where criterion features are rated present and the diagnostic concept to which they belong is present) are variable across PDs, with criterion features associated with AVD and DPS PDs being among the most sensitive and those associated with COM, PAR, and SZT among the least sensitive; and (4) that PPP values associated with criterion features across PDs tends to be low, indicating that criterion features are often rated as present among persons who are not true-positive cases for the diagnosis.

Stepwise Logistic Regression Analyses of PD Criterion Features

Table 5 also presents summaries of stepwise logistic regression analyses that evaluated the extent to which criteria within PD categories were uniquely able to discriminate positive from negative diagnoses. A predictor (i.e., a criterion feature of a PD category) was entered in a step if its presence improved the overall model's fit with the observed classification results (i.e., whether the PD category was diagnosed present or absent) as indexed by the significance of the increase in the likelihood ratio. Of the 63 PD criteria across eight PD categories that were evaluated in the regression analyses, the majority (70%) discriminated presence versus absence of diagnosis at a significance level of less than .05. AVD and DPS required the fewest number of criteria to predict diagnosis, an anticipated finding given the apparent psychometric superiority of the criteria associated with these concepts relative to others. Conversely, COM, psychometrically the weakest PD concept, required as many steps as there were criteria to maximally predict diagnosis. Overall, consideration of findings from these analyses along with the associated diagnostic efficiency statistics also presented in Table 5 and internal consistency values presented

in Table 4 suggest criterion features that can be eliminated because they contribute little useful or unique information to the diagnostic process. Elimination of such criteria would also function to narrow diagnostic boundaries and increase homogeneity among members of a PD diagnostic category.

DISCUSSION

The primary objectives of this research were to evaluate the psychometric and diagnostic efficiency indicators associated with PD criteria as assessed by the SCID-II. This research was conducted with a community-based sample who were psychotropic medication-free at the time of evaluation so as to eliminate the effect or influence that such medications may have on symptom covariation. Fifty-eight percent ($n = 86$) of 149 study participants were found to be positive for diagnosis of at least one PD as determined by the SCID-II, and among these, the average number of PDs observed was 2.22, a finding similar to that of another study that used similar recruitment strategies.⁴⁰

Consistent with its intended use as a screening device, several analyses suggest that the SCID-II-PQ is not a substitute for the SCID-II for determining the presence versus absence of specific DSM-IV PD diagnoses or the degree to which individuals display specific PD pathology. Findings from this research and other studies^{19,20,46} demonstrate that the SCID-II-PQ generally produces a high rate of false-positive PD diagnoses relative to SCID-II findings (67%). Similarly, kappa coefficients that indexed the degree of agreement in positive and negative diagnostic judgments derived by both measures tended to be low across PDs (median $\kappa = .37$). Although the SCID-II-PQ does better in predicting true-positive diagnoses for any PD based on SCID-II findings (67%), this study along with others⁷ suggests that self-report questionnaires do not result in diagnostic judgments that are comparable to those derived from semistructured interviews, and should probably not be used to infer the presence of specific DSM PD diagnoses. It is noteworthy, however, that SCID-II-PQ findings in this study predicted extremely well the positive presence of any erratic-dramatic cluster diagnosis. In each instance where the SCID-II-PQ suggested a positive erratic-dramatic diagnosis, a true-positive diagnosis within the cluster was observed based on SCID-II assess-

ment. Accordingly, the false-positive rate of SCID-II-PQ erratic-dramatic cluster diagnosis relative to the SCID-II was 0%, indicating that the SCID-II-PQ may perform well as a screening tool for PD diagnoses at the cluster level of analyses for erratic-dramatic PDs, but not for specific PD diagnoses within the cluster.

Findings from the SCID-II suggest generally modest levels of both comorbidity and covariation across PD concepts. Exceptions to this trend are BRD and PAG, and to a lesser extent, PAR and DPS, which tended to demonstrate greater comorbidity and/or covariation relative to other PD concepts. Features of this sample, namely, that it was community-based and consisted of persons who were psychotropic medication-free, may have resulted in somewhat attenuated patterns of comorbidity and covariation when compared with other studies. For example, among individuals with any positive PD in this study, the average number of PDs observed was 2.22. However, in studies that have used patient samples, an average of about four PD diagnoses tend to be made for individuals who are positive for at least one PD.⁴⁴ Comorbidity is also influenced by the prevalence rate of PDs within samples,⁴⁴ and in this particular sample, there were four PDs that occurred among 10% or less of the sample (DEP, SZD, HST, and ANT). Although the most representative evaluation of PD comorbidity and covariation patterns could only be established through random selection of members within the population, patterns observed within the current sample may more accurately reflect those of general population than studies that have used patient samples and/or samples that contained disproportionately high percentages of persons on psychotropic medications (e.g., inpatient samples) relative to the general population.

In terms of the criteria associated with specific PDs as assessed by the SCID-II, there is considerable variation in psychometric strength and diagnostic efficiency. The weakest PD concept in terms of these indices is COM. As a disorder, COM was reliably assessed across interviewers at both the level of diagnostic agreement ($\kappa = .908$) and at the level of symptom ratings (median $\kappa = .900$, ICC for summary dimensional scores for COM = .973). Consequently, unreliability of assessment among interviewers does not seem to explain the relatively poor showing of this PD concept. COM

criteria as assessed by the SCID-II do, however, demonstrate low internal consistency. This finding suggests that the COM concept is essentially a chimera, or comprised of a collection of weakly related features that have been melded together to form a diagnostic concept. Relatedly, COM criteria as assessed by the SCID-II had, on average, the lowest SEN and NPP values, and among the lowest SPC values. The mean PPP value across COM criteria, however, was the second highest among the PDs for which these analyses were performed, a finding that is probably in part a byproduct of the relatively high incidence of this diagnosis within the sample (22.1% as determined by the SCID-II). A logistic regression analysis suggested that all COM criteria need to be considered to maximize predictive accuracy associated with the diagnosis of disorder presence or absence, with this being the only PD that required sequential entry of all criteria into the predictive model. In interpreting these findings, however, some degree of caution must be exercised as it is impossible given the methods employed in this study to determine whether the COM PD concept performed poorly due to problems associated with SCID-II-PQ or SCID-II assessment or if the problems are more generally related to the PD category. For example, when SCID-II-PQ findings are referenced to those from the SCID-II, COM relative to other PDs has the highest false-positive rate (49%), even though it was the most commonly diagnosed PD from the SCID-II in this sample. These latter findings could suggest either problems with SCID-II-PQ items that assess COM criteria or problems with the PD criteria themselves. It is noteworthy that COM traditionally has performed differently from other PDs. For example, Nakao et al.⁴⁷ found less functional impairment associated with COM than other PDs. On the Millon Clinical Multiaxial Inventory-III,⁴⁸ the compulsive scale is consistently negatively correlated with a number of indices of psychopathology, including scale scores corresponding to other PD concepts (except for the histrionic and narcissistic scales, which also show consistent negative correlations with psychopathology measures). Some previous studies where PD dimensional scores were factor-analyzed revealed that COM loaded most highly on a factor that was defined by COM alone.⁴⁹⁻⁵¹ Given these findings, COM as a PD concept may well benefit

from additional conceptual, theoretical, and empirical development.

DEP very infrequently occurred in the sample ($n = 1$, 0.7%) as assessed by the SCID-II. Similarly, relatively few of the 149 participants ($n = 9$, 6%) endorsed enough items on the SCID-II-PQ to be considered for a DSM-IV DEP diagnosis. The DEP PD concept was considerably revised from DSM-III to DSM-III-R, and substantial changes were made in diagnostic criteria from DSM-III-R to DSM-IV, including a shift in the symptom cut point for diagnosis from four symptoms to five. It would appear that these changes may have resulted in a sharp decline in the prevalence of this condition from previous DSM editions, as pre-DSM-IV research has indicated substantial comorbidity of DEP with other PDs.⁵² Additionally, in this study, DEP criteria were relatively low in internal consistency. Weakly related criteria in combination with a relatively low base rate of criteria endorsement and a conservative symptom cut point for diagnosis likely resulted in the almost total absence of this diagnosis in the study.

Across several analyses, criteria for AVD performed relatively well. Among the PD criteria, they were the most internally consistent on SCID-II and consistently were among the strongest in terms of diagnostic efficiency. The provisional PD concepts of DPS also demonstrated strong internal consistency and relatively strong diagnostic efficiency.

From the analyses performed, several observations can be offered about the performance of individual PD criteria. These observations, however, must be considered with caution, as it is impossible to completely disentangle in this research potential problems associated with the assessment measures used from problems with the PD criteria themselves. As such, future research may seek to cross-validate these observations when PD criteria are assessed with other measures and different samples. One of the more consistent findings is the poor performance of criterion 5 for DPS (critical and judgmental towards others), which was found to detract from the overall internal consistency of the PD concept. Furthermore, this item had the weakest diagnostic efficiency statistics of the DPS criterion set. Criterion 7 of COM (miserly spending style) has similar characteristics, and appears to

detract from rather than enhance the definition of this diagnostic concept. Although criterion 4 (reads demeaning or threatening meanings into benign remarks or events) of PAR enhanced the internal consistency of the criterion set, it proved to be the weakest in terms of diagnostic efficiency and was the only criterion within the set that did not uniquely contribute to the accurate prediction of the PAR diagnosis. In considering these observations, one relevant issue is the inter-rater reliability of the criteria in question, as poor reliability may account for the relatively poor performances of these criteria. However, inter-rater reliability was adequate and does not appear to account for the poor performance of these criteria: DPS criterion 5 ($\kappa = .963$), COM criterion 7 ($\kappa = .775$), and PAR criterion 4 ($\kappa = .896$).

SZT criterion 4 (odd thinking or speech) and 6 (inappropriate or constricted affect) also performed weakly. These criteria are only assessed during the SCID-II based on interview impressions and observations. It may be the case that these criteria are substantially related to the SZT concept; however, the method of assessment (informal observations within a 1- to 2-hour block of time) may not be sufficient for assessing these criteria. It is also noteworthy that these observational items were among the least reliably assessed criteria in this study ($\kappa = .132$ for criterion 4; $\kappa = .110$ for criterion 6) and were substantially below the median kappa (.720) for agreement in ratings across the nine SZT criteria. Consequently, the observational approach to assessment associated with these criteria may be inherently unreliable. Reliability assessments were also performed from audiotapes and, consequently, important nonverbal cues relevant for observer judgments about these criteria may not have been conveyed through the sound recordings.

When results of the logistic regression analyses from this study are compared with those of Nurnberg et al.³⁷ based on DSM-III-R PD conceptualizations, some similarities and differences are apparent. For example, DSM-IV PAR criterion 2 (unjustified doubts about loyalty of friends and associates), SZT criterion 3 (unusual perceptual experiences), COM criterion 6 (stubborn insistence on his or her way of doing things), NAR criterion 5 (sense of entitlement),

and BRD criterion 7 (feelings of emptiness) emerged as consistently strong predictors across studies. However, DSM-IV PAR criterion 4 (reads hidden threatening meanings into benign events), SZT criterion 4 (odd thinking and speech), and BRD criterion 6 (affective instability) were unreliable in their ability to predict diagnosis. Such inconsistencies, however, may be the product of changes made in other diagnostic criteria within these PD concepts that, in turn, affected the overall likelihood of diagnosis from DSM-III-R to DSM-IV. Alternatively, these discrepancies may be the result in differences in measurement accuracy (i.e., reliability of assessment) or subject characteristics across studies. Overall, the logistic regression analyses suggest the need for additional replication-oriented research based on representative and reliably assessed samples of persons before firm conclusions can be offered as to which criteria should be retained to enhance diagnostic efficiency and homogeneity within PD groups.

A couple of methodological caveats must be considered in relation to this study's overall findings. First, participants were not screened for axis I disorders. There is evidence that the co-occurring presence of axis I conditions may alter the presentation of axis II features.⁵³ Given that the sample employed in this research was not a patient sample (although some participants were receiving psychosocial therapy at the time of study participation) and that all sample participants were psychotropic medication-free, one might speculate that the rate of co-occurring axis I disorders was less than would be typically observed in inpatient and outpatient settings. The possible effects of differences in axis I and II comorbidity in relation to this study's findings are uncertain and warrant future investigation. Second, the method used to recruit potential participants was relatively unique given methods commonly employed in the PD literature. Although the intent of such an approach was to enhance the ecological validity of PD assessment through the recruitment of persons beyond treatment settings, it is also clear that the obtained sample was different from the general population in important respects, most notably in the prevalence of PDs and PD pathology. Consequently, findings from the present research

should be contrasted with those subsequently obtained from population and patient samples.

DSM PD concepts have long been plagued by both conceptual and empirical problems,^{3,7} and have not been widely embraced by practitioners.^{5,6} The absence of a clearly stated rationale or theoretical basis for each PD concept coupled with the disappointing reliability and validity of PD assessment likely contribute to the various problems associated with DSM PDs. Alternative approaches to the conceptualization of PDs and the establishment of PD criteria for official nomenclatures should be considered. One such issue involves the level at which criterion features are defined. At present, DSM PD definitions are largely topographical, whereby criterion features are defined in terms of their descriptive or surface-level properties. Simultaneous consideration might also be given to the functions that covarying PD criterion features serve^{54,55} as a means of adding coherence and internal consistency to the various PD concepts. This proposal has some support in the present study. For example, AVD PD was among the PDs with the strongest psychometric properties. This disorder is defined by several criterion features that specify both topography (e.g., "avoids occupational activities that involve significant interpersonal contact. . .") and function ("...because of fears of criticism, disapproval, or rejection"). Another issue is that DSM PDs remain relatively divorced from the extensive theory and research on the dimensions of "normal" personality features (however, see Costa and Widiger⁵⁶ for examples of how PD conceptualization and theory can be influenced by or incorporated into well-researched and validated personality theories and models). In addition, the divergence between the methods used to develop PD criteria (committee consensus) and the methods used to develop and assess personality constructs (psychometric analysis, test and item response theories) likely contributes to the suboptimal reliability and validity of PD criteria. At minimum, greater emphasis needs to be placed on the empirical and statistical evaluation of prospective PD concepts. In conclusion, unless substantial changes are undertaken in terms of how PDs are conceptualized, specified, and evaluated, a clinical science of PDs is likely to progress very slowly.

REFERENCES

1. Blashfield RK, Intoccia V. Growth of the literature on the topic of personality disorders. *Am J Psychiatry* 2000;157:472-473.
2. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*. Ed. 3. Washington, DC: American Psychiatric Association, 1980.
3. Farmer RF. Issues in the assessment and conceptualization of personality disorders. *Clin Psychol Rev* 2000;20:823-851.
4. Livesley WJ. Suggestions for a framework for an empirically based classification of personality disorders. *Can J Psychiatry* 1998;43:137-147.
5. Maser JD, Kaelber C, Weise RE. International use and attitudes towards DSM-III and DSM-III-R: growing consensus in psychiatric classification. *J Abnorm Psychol* 1991;100:271-279.
6. Westen D, Arkowitz-Western L. Limitations of axis II in diagnosing personality pathology in clinical practice. *Am J Psychiatry* 1998;155:1767-1771.
7. Clark LA, Livesley WJ, Morey L. Personality disorder assessment: the challenge of construct validity. *J Pers Disord* 1997;11:205-231.
8. Zimmerman M. Diagnosing personality disorders: a review of issues and research methods. *Arch Gen Psychiatry* 1994;51:225-245.
9. Perry J. Problems and considerations in the valid assessment of personality disorders. *Am J Psychiatry* 1992;149:1645-1653.
10. Widiger TA, Sanderson CJ. Assessing personality disorders. In: Butcher JN (ed). *Clinical Personality Assessment: Practical Approaches*. New York, NY: Oxford University Press, 1995;380-394.
11. Modestin J, Erni T, Oberson B. A comparison of self-report and interview diagnoses of DSM-III-R personality disorders. *Eur J Pers* 1998;12:445-455.
12. Shea MT. Some characteristics of the axis II criteria sets and their implications for the assessment of personality disorders. *J Pers Disord* 1992;6:377-381.
13. Blais MA, Norman DK. A psychometric evaluation of the DSM-IV personality disorder criteria. *J Pers Disord* 1997;11:168-176.
14. Morey LC. A psychometric analysis of the DSM-III-R personality disorders criteria. *J Pers Disord* 1988;2:109-124.
15. Livesley WJ (ed). *The DSM-IV Personality Disorders*. New York, NY: Guilford, 1995.
16. First MB, Gibbon M, Spitzer RL, Williams JBW, Benjamin LS. *User's Guide for the Structured Clinical Interview for DSM-IV Axis II Personality Disorders (SCID-II)*. Washington, DC: American Psychiatric Press, 1997.
17. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*. Ed. 4. Washington, DC: American Psychiatric Association, 1994.
18. First MB, Spitzer RL, Gibbon M, Williams JBW. *The Structured Clinical Interview for DSM-III-R Personality Disorders (SCID-II)*. Part I: description. *J Pers Disord* 1995;9:83-91.
19. Nussbaum D, Rogers R. Screening psychiatric patients for Axis II disorders. *Can J Psychiatry* 1992;37:658-660.
20. Jacobsberg L, Perry S, Frances A. Diagnostic agreement between the SCID-II screening questionnaire and the Personality Disorder Examination. *J Pers Assess* 1995;65:428-433.
21. Brooks RB, Baltazar PL, McDowell DE, Munjack DJ, Bruns JR. Personality disorders co-occurring with panic disorder with agoraphobia. *J Pers Disord* 1991;5:328-336.
22. Dreessen L, Arntz A. Short-interval test-retest interrater reliability of the Structured Clinical Interview for DSM-III-R Personality Disorders (SCID-II) in outpatients. *J Pers Disord* 1998;12:138-148.
23. Fogelson DL, Neuchterlein KH, Asarnow RF, Subotnik KL, Talovic SA. Interrater reliability of the Structured Clinical Interview for DSM-III-R, axis II: schizophrenia spectrum and affective spectrum disorders. *Psychiatry Res* 1991;39:55-63.
24. Renneberg B, Chambless DL, Dowdall DJ, Fauerbach JA, Gracely EJ. The Structured Clinical Interview for DSM-III-R, axis II and the Millon Clinical Multiaxial Inventory: a concurrent validity study of personality disorders among anxious patients. *J Pers Disord* 1992;6:117-124.
25. Maffei C, Fossati A, Agostoni I, Barraco A, Bagnato M, Deborah D, et al. Interrater reliability and internal consistency of the Structured Clinical Interview for DSM-IV axis II Personality Disorders (SCID-II), Version 2.0. *J Pers Disord* 1997;11:279-284.
26. First MB, Spitzer RL, Gibbon M, Williams JBW, Davies M, Borus J, et al. *The Structured Clinical Interview for DSM-III-R Personality Disorders (SCID-II)*. Part II: multi-site test-retest reliability study. *J Pers Disord* 1995;9:92-104.
27. Weiss RD, Najavits LM, Muenz LR, Hufford C. Twelve-month test-retest reliability of the Structured Clinical Interview for DSM-III-R Personality Disorders in cocaine-dependent patients. *Compr Psychiatry* 1995;36:384-389.
28. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*. Ed. 3, Rev. Washington, DC: American Psychiatric Association, 1987.
29. Skodol AE, Rosnick L, Kellman D, Oldham JM, Hyler SE. Validating structured DSM-III-R personality disorder assessments with longitudinal data. *Am J Psychiatry* 1988;145:1297-1299.
30. Spitzer RL. Psychiatric diagnoses: are clinicians still necessary? *Compr Psychiatry* 1983;24:399-411.
31. Hyler SE, Skodol AE, Kellman HD, Oldham JM, Rosnick L. Validity of the Personality Diagnostic Questionnaire-Revised: comparison with two structured interviews. *Am J Psychiatry* 1990;147:1043-1048.
32. Hyler SE, Skodol AE, Oldham JM, Kellman HD, Doidege N. Validity of the Personality Diagnostic Questionnaire-Revised: a replication in an outpatient sample. *Compr Psychiatry* 1992;33:73-77.
33. O'Boyle M, Self D. A comparison of two interviews for DSM-III-R personality disorders. *Psychiatry Res* 1990;32:85-92.
34. Skodol AE, Oldham J, Rosnick L, Kellman HD, Hyler SE. Diagnosis of DSM-III-R personality disorders: A comparison with two structured interviews. *Intl J Methods Psychiatr Res* 1991;1:13-26.
35. Skodol AE, Rosnick L, Kellman HD, Oldham JM, Hyler S. Development of a procedure for validating structured assessments of axis II. In: Oldham JM (ed). *Personality Disorders: New Perspectives on Diagnostic Validity*. Washington, DC: American Psychiatric Press, 1991:43-70.
36. Barber JP, Morse JQ. Validation of the Wisconsin Per-

- sonality Disorders Inventory with the SCID-II and PDE. *J Pers Disord* 1994;8:307-319.
37. Nurnberg HG, Martin GA, Pollack S. An empirical method to refine personality disorder classification using stepwise logistic regression modeling to develop diagnostic criteria and thresholds. *Compr Psychiatry* 1994;35:409-419.
 38. Spitzer RL, Williams JBW, Gibbon M, First MB. *SCID Users Guide for the Structured Clinical Interview for DSM-III-R*. Washington, DC: American Psychiatric Press, 1990.
 39. Newman JP, Widom CS, Nathan S. Passive avoidance in syndromes of disinhibition: psychopathy and extraversion. *J Pers Soc Psychol* 1985;48:1316-1327.
 40. Farmer RF, Nelson-Gray RO. Anxiety, impulsivity, and the anxious-fearful and erratic-dramatic personality disorders. *J Res Pers* 1995;29:189-207.
 41. Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. *Psychol Bull* 1979;86:420-428.
 42. Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educ Psychol Measure* 1973;33:613-619.
 43. Fleiss JL. *Statistical Methods for Rates and Proportions*. Ed. 2. New York, NY: Wiley, 1981.
 44. Widiger TA, Frances A, Harris M, Jacobsberg L, Fyer M, Manning D. Comorbidity among axis II disorders. In: Oldham, JM (ed). *Personality Disorders: New Perspectives on Diagnostic Validity* Washington, DC: American Psychiatric Press, 1991: 165-194.
 45. Baldessarini RJ, Finklestein S, Arana GW. The predictive power of diagnostic tests and the effect of prevalence of illness. *Arch Gen Psychiatry* 1983;40:569-573.
 46. Ekselius L, Lindstrom E, von Knorring L, Bodlund O, Kullgren G. SCID II interviews and the SCID screen questionnaire as diagnostic tools for personality disorders in DSM-III-R. *Acta Psychiatr Scand* 1994;90:120-123.
 47. Nakao K, Gunderson JG, Phillips KA, Tanaka N, Yori-fugi K, Takaishi J, et al. Functional impairment in personality disorders. *J Pers Disord* 1992;6:24-33.
 48. Millon T. *MCMI-III Manual*. Minneapolis, MN: National Computer Systems, 1994.
 49. Hyler SE, Lyons SE. Factor structure of the DSM-III personality disorder clusters: a replication. *Compr Psychiatry* 1988;29:304-308.
 50. Kass F, Skodol A, Charles E, Spitzer R, Williams J. Scaled ratings of DSM-III personality disorders. *Am J Psychiatry* 1985;142:627-630.
 51. Livesley WJ, Jackson DN. The internal consistency and factorial structure of behaviors judged to be associated with DSM-III personality disorders. *Am J Psychiatry* 1986;143: 1473-1474.
 52. Hirschfeld RMA, Shea MT, Weise R. Dependent personality disorder. In: Livesley WJ (ed). *The DSM-IV Personality Disorders*. New York, NY: Guilford, 1993:239-256.
 53. Farmer RF, Nelson-Gray RO. Personality disorders and depression; hypothetical relations, empirical findings, and methodological considerations. *Clin Psychol Rev* 1990;10:453-476.
 54. Nelson-Gray RO, Farmer RF. Behavioral assessment of personality disorders. *Behav Res Ther* 1999;37:347-368.
 55. Farmer RF, Nelson-Gray RO. Functional analysis and response covariation in the assessment of personality disorders. A reply to Staats and to Bissett and Hayes. *Behav Res Ther* 1999;37:385-394.
 56. Costa PT Jr, Widiger TA (eds). *Personality Disorders and the Five-Factor Model of Personality*. New York, NY: Guilford, 1994.