Journal of Personality Disorders, 17(4), 351-369, 2003

ASSESSING PERSONALITY DISORDERS USING A SYSTEMATIC CLINICAL INTERVIEW: EVALUATION OF AN ALTERNATIVE TO STRUCTURED INTERVIEWS

Drew Westen, PhD, and Serra Muderrisoglu, PhD

The aim of this study was to assess interrater reliability and provide initial data bearing on the validity of a method of assessing personality disorders (PDs) that does not presume that patients can accurately self-report personality pathology. In a sample of 24 outpatients, two clinician-judges independently applied the Shedler-Westen Assessment Procedure-200 (SWAP-200; Westen & Shedler, 1999a, 2000), a 200-item Q-sort procedure for assessing personality pathology, to data from the Clinical Diagnostic Interview (Westen, 2002), a systematic clinical interview that mirrors and standardizes methods used by experienced clinicians to diagnose personality. In 16 of the 24 cases, the treating clinician also independently described the patient using the SWAP-200 Q-sort, based on longitudinal knowledge of the patient over the course of treatment, blind to the interview data. Interrater reliability was uniformly high, with median correlations between interviewers at $r > .80$. Interviewer-treating clinician correlations were also high, with median convergent validity coefficients at $r > .80$. Diagnostic overlap (discriminant validity) was moderate for dimensional DSM-IV diagnoses, reflecting extensive comorbidity among disorders, but minimal for empirically derived diagnoses identified in prior research. Treating clinicians' dimensional PD diagnoses using this method also strongly predicted interviewer-rated measures of adaptive functioning. The findings provide preliminary support for the reliability and validity of an alternative to structured interviews for diagnosing personality pathology, and suggest that the way to improve validity of personality diagnosis may not be to minimize clinical inference but to quantify it using psychometric instruments.

The development of structured interviews to assess personality disorders (PDs) has made possible tremendous advances in the field since their emer-

From the Departments of Psychology and Psychiatry and Behavioral Sciences, Emory University (D. W.), and the Department of Psychology, Bogazici University, Istanbul, Turkey (S. M.).

Preparation of this manuscript was supported in part by NIMH MH59685 and MH60892 to the first author, and by a grant from the Milton Fund at Harvard Medical School.

The authors thank Samantha Glass and Drs. Christopher Fowler, Amy Kegley Heim, Heather Thompson, and Carolyn Zittel for their contributions to this study.

Address correspondence to Drew Westen, PhD, Department of Psychology, Emory University, 532 N. Kilgo Cir., Atlanta, GA 30322; E-mail: dwesten@emory.edu.

gence around the time of DSM-III (First et al., 1995; Gunderson, Kolb, & Austin, 1981; Loranger, Susman, Oldham, & Russakoff, 1998; Pfohl, Stangl, Zimmerman, Bowers, & Corenthal, 1995; Zanarini, Frankenberg, Chauncey, & Gunderson, 1987). Nevertheless, structured interview methods face particular challenges when applied to PDs, reflected in their inconsistent associations with other instruments assessing the same constructs or with consensus clinical diagnoses using all available data (see Perry, 1992; Pilkonis, Heape, Ruddy, & Serrao,1991, 1995; Pilkonis et al., 1995; Westen, 1997). Although some studies have produced median kappas indexing diagnostic convergence across two structured interviews as high as $\kappa = .45$ (SCID-II and PDE for DSM-III-R; Skodol, Oldham, Rosnick, Kellman, & Hyler, 1991), across studies the median kappa tends to be closer to .30, and the median *r* assessing convergence of dimensional diagnoses is in the range of $r = .40$ (see Clark, Livesley, & Morey, 1997). For some disorders, such as antisocial PD, cross-instrument validity coefficients are high, whereas for disorders whose symptoms are less overt (e.g., narcissistic, or passive-aggressive, whose deletion from DSM-IV was influenced by the difficulty in measuring it reliably), validity tends to be considerably weaker, with many studies showing little or no association between the same instrument and another instrument putatively assessing the same construct, often using similar questions.

In the absence of a gold standard against which to compare PD diagnoses using structured interviews, some researchers have turned to a LEAD standard (longitudinal expert evaluation using all available data) (Spitzer, 1983). To make a LEAD diagnosis, multiple members of an investigatory team with knowledge of the patient from different sources and at different times meet to arrive at a consensus diagnosis. Although this method, like all others, is certainly fallible, its advantage is that it assesses behaviors and traits that are by definition characteristic of the patient's functioning over time (not just on one occasion), present in various contexts (which are most likely to be seen in different situations), and consistent despite state changes (e.g., depressed mood) that can influence responses to questions on any given occasion (Skodol et al., 1991).

Studies comparing clinical diagnoses using the LEAD standard and structured interviews generally show only weak concordance. Skodol and colleagues (1991) found that the two most widely used PD instruments (the PDE and the SCID-II) showed moderate convergence with each other, with a range of kappas from .14 to .66; however, concordance with the LEAD standard was low for both instruments. For the SCID-II, kappas ranged from .03 (narcissistic and passive-aggressive) to .60 (dependent); for the PDE kappas ranged from -.01 (passive-aggressive) to .41 (dependent); and for both instruments the median kappa was $\kappa = .25$. In another study, neither the PDE nor the SIDP converged with consensus clinical diagnoses using all available data (including systematic informant interviews) in a sample of 108 patients (Pilkonis et al., 1995). For example, concordance with the LEAD standard on diagnosis of any-PD versus no-PD yielded kappas of only .18 for the PDE and .37 for the SIDP; individual diagnoses tended to fare no better. Research on the association between LEAD diagnoses and self-report ques-

tionnaires, though limited, shows even less convergence. For example, Wilberg, Dammen, & Friis (2000) examined the relation between the PDQ-4 questionnaire and the LEAD standard. Kappas ranged from $\kappa = .05$ to .26, with a median $\kappa = .13$.

Validity evidence on structured interviews for Axis I and Axis II disorders differs substantially in this respect. Diagnoses made using structured interviews for Axis I disorders (e.g., SCID-diagnosed Major Depressive Disorder) tend to show strong associations with other interview measures (e.g., the Hamilton Rating Scale for Depression; Hamilton, 1960) as well as with self-report questionaires (e.g., the Beck Depression Inventory). In contrast, PD interviews show moderate to low associations with other PD interviews and particularly low correlations with self-report PD measures (e.g., Hills, 1995). As we have argued elsewhere (Westen, 1997; Westen & Arkowitz-Westen, 1998), the difference may be best understood in its historical context. The success of structured interviews for disorders such as major depression in the late 1970s led researchers to import a measurement strategy designed for Axis I syndromes to the assessment of PDs, namely asking questions derived from DSM criteria. Clinicians of all theoretical perspectives report that they do not diagnose personality primarily by asking such questions (although they certainly do not discount direct answers); they rely, instead, primarily on the narratives patients provide in describing emotionally significant events, and secondarily on the patient's behavior in the consulting room. Interestingly, researchers have come to similar conclusions in the assessment of adult attachment patterns, finding only minimal convergence between self-reported attachment styles and reliably-rated (and strongly predictive) attachment styles coded from narratives (e.g., Cassidy & Mohr, 2002; Main, Kaplan, & Cassidy, 1985).

THE PRESENT STUDY

In the present study, we applied the Shedler-Westen Assessment Procedure (SWAP-200), a Q-sort instrument designed to quantify the judgment of experienced clinical interviewers (Westen & Shedler, 1999a, 2000), to data from the Clinical Diagnostic Interview (CDI; see Westen, Muderrisoglu, Fowler, Shedler, & Koren, 1997). In contrast to a structured interview, the CDI is what might be called a *systematic clinical interview*, designed to mirror but systematize and standardize the kind of interviewing approach typically used by experienced clinicians. This approach differs in three respects from structured interviewing. First, although the CDI includes some direct questions (e.g., about characteristic moods, subclinical thinking disturbances), it does not primarily ask patients to describe their personalities. Rather, it asks them to tell narratives about their lives that allow the interviewer to make judgments about their characteristic ways of thinking, feeling, regulating emotions, experiencing themselves and others, and so forth. Second, the interview is not organized around the diagnostic criteria for each of the DSM-IV PDs (e.g., one question assessing whether the patient is entitled). Instead, after completing the interview, the interviewer sorts the 200 items of the SWAP-200 according to the degree to which they apply to the patient, relying not only on what the patient says but also the way he or

she says it. Third, rather than making diagnoses by counting symptoms, diagnosis using this method involves prototype matching, in which patients' diagnostic scores reflect the degree of match (correlation) between their 200-item profile and empirical prototypes of each diagnosis. The resulting scores (converted to T-scores, with a mean of 50 and *sd* of 10) can be treated dimensionally or categorically (by selecting cutoffs, as in an MMPI-2 profile).

The study had three aims. The first two aims addressed questions of reliability and validity of personality diagnosis using the SWAP-200. To date, the published literature on the SWAP-200 has relied almost exclusively on data from large random samples of experienced doctoral-level clinicians, with unknown reliability (e.g., Westen & Harnden-Fischer, 2001; Westen & Shedler, 1999a, 1999b; Westen, Shedler, Durrett, Glass, & Martens, 2003). Thus, the first aim was to assess the interrater reliability of dimensional PD diagnoses made by interview, to see if this diagnostic method might be useful in psychiatric research (i.e., research that does not rely on clinicians as primary informants). The second aim was to examine the extent to which SWAP-200 diagnostic judgments made by treating clinicians resemble those made by interview, and to asses the extent to which these judgments predict measures of adaptive functioning bearing on validity of a PD instrument. Strong correlations between treating-clinician data and independent interview assessments would support the validity of clinician-report SWAP-200 data. The third aim was to assess the extent to which the high degree of diagnostic overlap (comorbidity) seen among PDs in studies using structured interviews is inherent in assessment of PDs or is an artifact of the categories and criteria included on Axis II.

METHODS

PARTICIPANTS

Participants were 24 patients from the Outpatient Psychiatry Department of The Cambridge Hospital or from the private practice of clinicians affiliated with the hospital or the research group willing to refer patients to the study. Clinicians were recruited by sending out a memorandum to the outpatient department requesting clinicians willing to refer a patient to the study and to complete a Q-sort description of the patient themselves. The only exclusion criterion was a history of psychosis. All patients were in individual psychotherapy, with no limits placed on length of treatment, which ranged from several weeks to several years. Treating clinicians were psychologists and psychiatrists who ranged from advanced residents to senior faculty with over 2 decades of experience. Clinicians willing to participate informed potential patients of the study and asked if they would give their permission to be contacted by the researchers. The researchers then met with potential subjects to discuss the study and obtain informed consent; 24 of 26 patients who contacted the researchers agreed to participate. Patients were paid $25 for their participation. Of the 24 clinicians who recruited patients, 16 provided a SWAP-description of the patient, for which they received an honorarium of $35.

PROCEDURES

Patients were interviewed by one of the authors or by a third interviewer, an advanced clinical psychology fellow trained in the use of the interview, using the CDI. Interviews were videotaped so that a second clinician-judge could evaluate the patient. Thus, 2 clinician-judges provided Q-sort descriptions of all 24 patients using the SWAP-200 (1 who conducted the interview, and the other who coded it from the videotape), blind to all data, including clinician diagnosis and each other's Q-sort descriptions of the patient. Participants were outpatients; hence, clinician-judges had no prior clinical interactions with them. The treating clinicians independently provided a Q-sort description based on their clinical experience with the patient over months or years. Treating clinicians were blind to all interview data.

The CDI asks patients to provide narratives about themselves, their lives, and their problems. It begins, as in a standard clinical interview, by asking what brought them in for treatment, with the interviewer probing for details about severity, frequency, duration, and history of symptoms and concerns. The interviewer then asks patients about a wide range of significant relationships and experiences from the past and present (e.g., parents, siblings, romantic relationships, friendships, school and work experiences), about particularly stressful or difficult times in their recent lives (to obtain information about how the patient appraises and copes with difficult circumstances), about their moods and emotions, and about their characteristic ways of thinking (to obtain data on subclinical thinking disturbances). For each of these categories of relationships or experiences, the interviewer follows general questions (e.g., "Can you tell me about your relationship with your wife?") with the request to describe two to three specific incidents. The first time such incidents are requested, the interviewer asks the participant to be sure to describe what led up to the event, what both people were thinking and feeling, and the outcome. The interview assumes a competent, experienced clinical interviewer familiar with the Q-sort items. Thus, although some sections of the interview suggest specific probes, most probing depends on clinical judgment and knowledge of what the interviewer needs to know in order to provide a valid Q-sort description of the patient following the interview.

MEASURES

Shedler-Westen Assessment Procedure-200. The SWAP-200 (Westen & Shedler, 1999a, 1999b, 2000) is a Q-sort designed to assess personality and personality pathology. A Q-Sort is a set of statements printed on separate index cards, in this case, statements about personality and personality dysfunction. In the present implementation of the Q-sort method (the SWAP-200), an experienced clinician sorts (rank-orders) the statements into categories (piles), from those that are not descriptive of the patient (assigned a value of "0") to those that are highly descriptive (assigned a value of "7"), with intermediate places of items that apply to varying degrees. Clinician-judges can sort the items using a fixed distribution (see Block, 1978),

based on data from the Clinical Diagnostic Interview or from their knowledge of the patient over the course of treatment (as in LEAD diagnosis).

The SWAP-200 thus provides a numeric score ranging from 0 to 7 for each of 200 personality-descriptive items. Items were derived from a number of sources, including DSM-III-R and DSM-IV Axis II criteria, clinical and empirical literature on PDs, research on normal personality traits and psychological health, and pilot interviews. Development of the item set was an iterative process that took approximately 7 years, using standard psychometric methods such as asking hundreds of experienced clinicians to use the instrument over several iterations to describe a patient and comment on anything important about the patient's personality they could not describe using the item set (to maximize content validity), eliminating items that proved empirically redundant, eliminating or rewording items with minimal variance, etc.

Research thus far supports the validity and reliability of the instrument in predicting objective indicators of personality dysfunction such as suicide attempts and history of psychiatric hospitalizations; adaptive functioning, assessed by measures such as the GAF; clinician diagnoses; and developmental and family history variables (Shedler & Westen, 2002; Westen & Harnden-Fischer, 2001; Westen & Shedler, 1999a). For example, a recent study of the adolescent version of the instrument found an association between histrionic PD and history of sexual abuse and disrupted attachments in girls, and between histrionic PD and history of sexual abuse and family history of bipolar illness in boys (see Westen & Chang, 2000). The adolescent version of the instrument has also been shown to predict patterns of association with relevant scales from the Child Behavior Checklist (Dutra, Campbell, & Westen, in press) and measures of attachment status (Nakash-Eisikovits, Dierberger, & Westen, 2002).

Adaptive Functioning. To provide validity data other than correlations between treating-clinician and interviewer dimensional diagnoses, a set of judges (including the 2 authors and 2-5 advanced graduate students in clinical psychology per subject) watched the videotapes and made ratings of adaptive functioning, including GAF ratings and 7-point Likert-type ratings of interpersonal functioning (1= very poor, 7= close and loving) and occupational functioning (1= unable to keep a job, 3 = unstable, 5 = stable, 7 = working to potential). We used these to assess the validity of independently-made clinician dimensional diagnoses using the SWAP-200.

STATISTICAL ANALYSIS

SWAP-200 PD scores (dimensional diagnoses) reflect the degree of match, or correlation, between patients' 200-item profiles and SWAP-200 prototypes of each disorder from a normative sample (Westen & Shedler, 1999a). For ease of interpretation, these correlations are converted to T-scores. Thus, a patient's dimensional score for each PD (which we will refer to hereafter as a *PD score*) reflects the degree of match between an observer's description of the patient (either an interviewer or the treating clinician) and an empirical prototype. Elsewhere we have shown that these scores correlate strongly with PD diagnoses using DSM-IV criteria in both adults and

adolescents, as well as with external measures of etiology and adaptive functioning (Westen & Chang, 2000; Westen & Shedler, 1999a, 1999b; Westen et al., 2003). The use of Q-correlations of this sort is not novel, and has a number of useful psychometric properties (see Block, 1978). Like the approach used in the MMPI-2, SWAP-200 PD scores reflect an empirical criterion-keying approach, which diagnoses pathology based on a measure of match between the patient's profile and a normative sample of patients with a given diagnosis used as a criterion group. However, unlike both the DSM-IV and the MMPI-2, on which changing a small number of responses from "no" to "yes" can dramatically alter a scale score or diagnosis (because each scale or diagnosis has a relatively small number of items or criteria), this approach is not sensitive to minor fluctuations in item rankings. Because the prototype-matching algorithm we use (a simple correlation coefficient) takes into account the entire configuration of 200 items, a substantial shift in ranking of four or five items will typically have only a small effect on the magnitude of the correlation (unless they are the most diagnostic descriptors of the disorder, in which case their ranking *should* affect the diagnosis).

To address the first two aims, we conducted the following analyses. To assess interrater reliability, we correlated the independently obtained scores of the two interview judges. To assess convergence between interviewer and treating–clinician dimensional diagnoses, we correlated the means of each of the two interview judges' scores for each PD (averaging scores from the two judges to maximize reliability of measurement) with PD scores from the SWAP-200 independently obtained from the treating clinician. We included as well a psychological health index, constructed by correlating each patient's Q-sort profile from each observer with an aggregated prototype of a high-functioning patient from our normative study (again converting to T-scores for ease of interpretation). To provide initial data on validity of clinician diagnoses using a criterion variable other than diagnosis, we correlated clinicians' PD scores with independent ratings of adaptive functioning made by interview.

To assess the third aim (distinguishing comorbidity that may be inherent in PD diagnosis from comorbidity inherent in the particular categories included on Axis II in DSM-IV), we examined correlations across observers using an empirically derived classification of PDs comprised of seven primary diagnostic prototypes designed to identify nonoverlapping diagnostic groupings (Westen & Shedler, 1999b, 2000). The classification was constructed using a cluster-analytic procedure, Q-analysis, applied to SWAP-200 data from a sample of 496 patients described by a random national sample of clinicians. The seven primary diagnoses include *dysphoric* (an anxious-depressive personality configuration), *antisocial-psychopathic* (a blend of antisocial and classically psychopathic features, such as remorselessness and lack of empathy), *schizoid* (a diagnosis characterized by many features currently included in schizoid and schizotypal PDs), *paranoid* (similar to the current paranoid diagnosis), *histrionic* (characterized by classic histrionic features as well as features such as rejection-sensitivity and fears of aloneness and abandonment currently included in the borderline and dependent diagnoses), *obsessive* (a high-functioning obsessional

style, characterized by both a capacity to work productively and classic obsessive features), and *narcissistic* (similar to the current narcissistic diagnosis in DSM-IV).[1]

Dimensional PD diagnoses using this alternative classification are similarly constructed by correlating subjects' 200-item Q-sort profiles with empirically derived prototypes and converting them to T-scores. To see whether the nonredundancy of diagnoses in our derivation sample could be replicated across observers in the present study, we once again examined both correlations between the two interviewers (interrater reliability) and between treating clinicians' PD scores and interviewer PD scores (aggregated across the two interviewers). To assess the extent to which empirical derivation of diagnoses eliminates artifactual comorbidity introduced by the DSM classification system, for all analyses, we report the data first for DSM-IV diagnoses and then for the seven empirically derived PD diagnoses.

RESULTS

The sample consisted of 16 women and 8 men, ranging from age 19 to 57, with a mean of 38.6 (*sd* = 10.5). The group was varied diagnostically. Axis I diagnoses (made by one of us [D.W.] based on all available data) included major depressive disorder (*N* = 9), dysthymic disorder (*N* = 7), adjustment disorder (*N* = 3), panic disorder (*N* = 2), substance use disorder (*N* = 2), eating disorder NOS (*N* = 2), dissociative disorder NOS (*N* = 1), and posttraumatic stress disorder (*N* = 1). Axis II categorical diagnoses included PD NOS (*N* = 9), borderline PD (*N* = 3), dependent PD (*N* = 2), and antisocial, avoidant, narcissistic PD (*N* = 1 each). As assessed by dimensional SWAP-200 PD scores assessed by interview, mean T-scores ranged from 43.1 (obsessive-compulsive PD) to 54.5 (borderline PD), with standard deviations ranging from 8.2 (dependent PD) to 10.5 (borderline PD). All PDs showed elevations of at least 1 *sd* for at least one patient (using norms from our national normative sample). These data suggest that the sample included a wide range of Axis II pathology. Patients whose data were included in the subsample on which we had data from the treating clinician did not differ significantly on any demographic or diagnostic variable from patients for whom we had only interview data.

INTERRATER RELIABILITY

Tables 1 and 2 provide data on interrater reliability (Pearson's *r*). Table 1 reports interrater reliability of SWAP PD scores using DSM-IV PD diagnoses,

1. The Q-analysis also produced five dysphoric subtypes, including the following: avoidant (similar to the current avoidant diagnosis), neurotic depressive (a high functioning, depressive style characterized by considerable strengths as well as chronic problems with self-esteem and self-criticism), emotionally dysregulated (a diagnosis similar to the current borderline diagnosis, but without histrionic features, and empirically independent of the empirically derived histrionic diagnosis, unlike the current borderline diagnosis), dependent-victimized (a blend of current dependent and self-defeating features), and hostile-externalizing (similar to passive-aggressive PD from DSM-III-R, with an emphasis on hostility and denial of responsibility). For simplicity, in this article we will focus only on the seven primary diagnoses, although analyses using the dysphoric subtypes, available from the first author, produced similar data.

whereas Table 2 provides data on reliability using our empirically derived prototypes. As can be seen by examining the correlations along the diagonal in Table 1, interrater reliability for the DSM-IV PDs was uniformly high. Reliability coefficients off the diagonal were moderate, suggesting that, like other instruments for assessing DSM-IV PDs (for which such data are usually not reported, but when they are, they show poor discrimination among disorders; see Clark et al., 1997), the SWAP-200 cannot clearly discriminate between related PDs, either because of shortcomings of the method or because of comorbidity built into the current Axis II diagnoses.

Fortunately, we were able to distinguish between those two explanations, by examining reliability coefficients for the seven empirically derived PD prototypes (Table 2). The data address two questions. The first is whether we were able to make dimensional PD diagnoses reliably using these prototypes. The data along the diagonal show impressive interrater reliability, with a range of $r = .73$ to .87 and a median $r = .81$. Equally important, however, is a second question regarding what might be called "discriminant reliability" (the magnitude of correlations off the diagonal): Did the diagnostic overlap seen in Table 1 reflect problems of comorbidity specific to the DSM classification system or problems with our method of assessing personality pathology? The data from Table 2 demonstrate that the problem lies with the DSM, and that our assessment procedure is highly specific in the inferences it allows clinician-judges to draw reliably. Whereas the median correlation along the diagonal was $r = .81$, the median correlation off the diagonal was $r = -.04$, and the median absolute value of correlations off the diagonal (which is particularly conservative, because it treats the high baseline intercorrelations of disorders seen in PD research since DSM-III no differently from negative correlations that are in fact discriminating) was $r = .19$. Although high interrater reliability is not uncommon in PD research, we are not aware of any study of PDs showing comparable cross-observer discriminations.

CORRELATIONS BETWEEN INTERVIEW AND TREATING-CLINICIAN DIAGNOSES

Tables 3 and 4 show the correlations between SWAP-200 PD scores assessed by interview and by the treating clinician. Table 3 reports data for DSM-IV PD diagnoses. As can be seen by examining the correlations along the diagonal, convergent validity coefficients are very high by standards of research in personality, where correlations across observers in the range of $r = .30$ to .50 are more typical. The magnitude of the correlations is also substantially larger than correlations between structured interviews and LEAD diagnoses (which are more reliable than our clinician diagnoses, which favors studies using consensus LEAD diagnosis over judgments made by a single clinician). The correlations along the diagonal range from $r = .55$ to .86, with a median $r = .74$. Discriminant validity is modest, again reflecting substantial overlap among DSM-IV diagnoses. Table 4 reports correlations between dimensional diagnoses by interview and by the treating clinician using our seven empirically derived PD prototypes. The coefficients along

TABLE 1. Interrater Reliability for Dimensional DSM–IV Personality Disorder Diagnoses (*N* = 24)

	Interview Judge 1									
Interview Judge 1	**Paranoid**	**Schizoid**	**Schizotypal**	**Antisocial**	**Borderline**	**Histrionic**	**Narcissistic**	**Avoidant**	**Dependent**	**Obsessive–Compulsive**
Paranoid	**.76+**	.08	.35	.67+	.64+	.39	.62+	–.12	–.18	–.27
Schizoid	.14	**.91+**	.65+	–.28	–.17	–.45*	–.36	.83+	.51*	.56**
Schizotypal	.32	.72+	**.86+**	.04	.34	.07	–.16	.49*	.42*	.03
Antisocial	.57**	–.28	.05	**.86+**	.59**	.52**	.69+	–.48*	–.48*	–.53+
Borderline	.26	–.22	.16	.44*	**.83+**	.60**	.27	–.27	–.42	–.64+
Histrionic	.18	–.36	.05	.44*	.59**	**.79+**	.44*	–.50*	–.02	–.68+
Narcissistic	.62+	–.32	–.03	.75+	.47*	.51*	**.83+**	–.50*	–.43*	–.38
Avoidant	–.16	.69+	.27	–.45*	–.37	–.53**	–.46*	**.85+**	.60**	.59**
Dependent	–.44*	.34	–.01	–.61**	–.32	–.27	–.48*	.63+	**.81+**	.35
Obsessive–Compulsive	–.14	.51**	–.02	–.52**	–.63+	–.78+	–.38	.70+	.33	**.84+**

Note. +Correlation is significant at the .001 level (2–tailed). **Correlation is significant at the 0.01 level (2–tailed). *Correlation is significant at the 0.05 level (2–tailed).

TABLE 2. Interrater Reliability for Empirically Derived Personality Disorder Diagnoses (*N* = 24)

	Interview Judge 2						
Interview Judge 1	**Dysphoric**	**Antisocial**	**Schizoid**	**Paranoid**	**Obsessional**	**Histrionic**	**Narcissistic**
Dysphoric	**.81+**	.–.37	.29	.–.42*	.28	–.21	.08
Antisocial	–.26	**.82+**	–.09	.38	–.54**	–.16	–.04
Schizoid	.54**	.06	**.89+**	.06	–.12	–.27	–.18
Paranoid	–.14	.46*	–.02	**.72+**	–.61**	–.00	–.09
Obsessive	.05	–.37	–.09	–.25	**.89+**	–.19	.36
Histrionic	–.25	–.13	–.25	–.13	–.04	**.79+**	.03
Narcissistic	–.30	–.04	.06	.24	.21	–.05	**.77****

Note. +Correlation is significant at the .001 level (2–tailed). **Correlation is significant at the 0.01 level (2–tailed). *Correlation is significant at the 0.05 level (2–tailed).

TABLE 3. Correlations between Interview and Treating–Clinician Dimensional Diagnoses for the DSM–IV Personality Disorders (*N* = 16)

	Treating Clinician Scores									
Interview Mean Scores	**Paranoid**	**Schizoid**	**Schizotypal**	**Antisocial**	**Borderline**	**Histrionic**	**Narcissistic**	**Avoidant**	**Dependent**	**Obsessive–Compulsive**
Paranoid	**.55***	.06	.24	.58*	.58*	.34	.43	–.02	–.15	–.15
Schizoid	.16	**.68****	.54*	–.22	–.28	–.53*	–.30	.60*	.26	.64**
Schizotypal	.29	.43	**.62***	.14	.19	.03	.01	.27	.11	.13
Antisocial	.63**	–.07	.19	**.86+**	.63**	.60*	.71**	–.19	–.25	–.42
Borderline	.36	–.15	.20	.56*	**.82+**	.64**	.39	–.17	.02	–.56*
Histrionic	.15	–.36	.04	.44	.48	**.80+**	.46	–.44	–.09	–.75+
Narcissistic	.43	–.23	–.05	.58*	.50*	.62*	**.67****	–.28	–.22	–.43
Avoidant	–.01	.75+	.47	–.41	–.40	–.68**	–.51*	**.82+**	.59*	.73+
Dependent	–.24	.43	.36	–.48	–.35	–.34	–.52*	.50	**.67****	.30
Obsessive–Compulsive	–.16	.46	.02	–.58*	–.63**	–.86+	–.54*	.51*	.21	**.82+**

Note. +Correlation is significant at the .001 level (2–tailed); **Correlation is significant at the 0.01 level (2–tailed); *Correlation is significant at the 0.05 level (2–tailed).

the diagonal are once again very large, with a median $r = .82$. Discriminant validity is generally strong, with a median $r = -.04$ and median absolute value of correlations off the diagonal $r = .18$.

One potential objection to the findings thus far is that the magnitude of the convergent and discriminant coefficients reflects some psychometric artifact of the algorithm we used to assign patients scores, namely the use of correlation coefficients (between subjects' profiles and the profiles of criterion groups). This is unlikely on the face of it, because the probability that an artifact could consistently produce highly specific patterns of correlations, in which the coefficients along the diagonal are high and the others are comparatively low, is prohibitively small. Nevertheless, to rule out this alternative explanation, we performed the same analyses using SWAP-200 trait (rather than PD) scores derived using conventional factor analysis (rather than Q-factor analysis) and found similar patterns of convergent and discriminant reliability and validity for factor-based scores.

PREDICTING ADAPTIVE FUNCTIONING

In a final set of analyses, we examined the relation between treating-clinician SWAP-200 diagnoses and measures of adaptive functioning assessed from the Clinical Diagnostic Interview, such as Global Assessment of Functioning (GAF) scores and 1 to 7 ratings of social and occupational functioning (mean interrater reliability of all observer pairs on all 3 measures of adaptive functioning $r > .70$). Table 5 reports the correlations between measures of adaptive functioning assessed by interview and SWAP-200 PD scores assessed by the treating clinician. We also include data on the SWAP-200 psychological health index, an index of healthy personality functioning. (We report here only the correlations between adaptive functioning variables assessed from the interviews and treating-clinician SWAP-200 PD scores because the data are completely independent; we obtained similar results when we correlated interview-based PD scores with clinician-rated adaptive functioning.)[2]

As can be seen from Table 5, the data support the validity of the SWAP-200, linking clinician diagnosis to objective measures of functioning assessed independently by interview. The data also suggest, as have the findings of many other studies, that the current DSM-IV disorders form a hierarchy of dysfunction. The three Cluster A disorders (paranoid, schizoid, and schizotypal) appear, along with two Cluster B disorders (borderline and antisocial), to be the most severe of the PDs, as indicated by high negative correlations with all three measures of adaptive functioning. (These data suggest that the increasing trend in PD research to analyze data at the level of DSM-IV clusters rather than individual diagnoses, motivated by problems of comorbidity, may be especially problematic for the Cluster B disorders, which vary substantially in degree of severity.) In contrast, the

2. We obtained similar correlations when we used the empirically derived rather than the DSM-IV, diagnoses, although the empirically derived histrionic, narcissistic, and obsessional diagnoses all appeared somewhat less disturbed than their DSM-IV counterparts.

TABLE 4. Correlations between Interview and Treating–Clinician Dimensional Diagnoses for the Empirically Derived Personality Disorders (*N* = 16)

	Treating Clinician Scores						
Interview Mean Scores	**Dysphoric**	**Antisocial**	**Schizoid**	**Paranoid**	**Obsessional**	**Histrionic**	**Narcissistic**
Dysphoric	**.82+**	.–.10	.58*	.–.32*	.04	–.51*	–.05
Antisocial	–.03	**.83+**	.15	.49	–.52*	.02	–.10
Schizoid	.29	.05	**.66****	–.02	–.04	–.36	.13
Paranoid	–.02	.31	–.27	**.53***	–.36	.04	–.18
Obsessive	–.08	–.54*	–.19	–.27	**.89+**	–.22	.20
Histrionic	–.52*	–.29	–.28	–.16	.00	**.82+**	.08
Narcissistic	.23	–.51*	.10	–.18	.32	.01	**.80****

Note. +Correlation is significant at the .001 level (2–tailed); **Correlation is significant at the 0.01 level (2–tailed); *Correlation is significant at the 0.05 level (2–tailed).

obsessive-compulsive prototype was slightly positively correlated with adaptive functioning, reflecting the fact that this prototype includes a number of items indexing adaptive traits such as conscientiousness. In fact, in many studies obsessive-compulsive PD has been an outlier that neither loads on factors with other PDs nor correlates negatively with measures of adaptive functioning (see, e.g., Pfohl & Blum, 1995). As the table also shows, the psychological health index derived from the SWAP-200 profile provided by the treating clinician strongly predicted every measure of adaptive functioning, suggesting that the instrument can assess degree of personality health/sickness independent of diagnosis.

DISCUSSION

LIMITATIONS

This study has two primary limitations. The first and most important is the limited sample size. A critic might argue that the sample is too small to draw meaningful conclusions. The data presented here clearly warrant further investigation with a substantially larger sample. It is important to note, however, that not only are the effect sizes reported here extremely large by any standards (e.g., median correlations between treating clinician and interview diagnoses of .80), but findings significant at $p < .001$ with a small sample have precisely the same meaning as findings at the same probability level obtained with larger *N*s. Perhaps most importantly, we were not predicting correlations taken one at a time; we were predicting that a set of convergent and discriminant reliability and validity matrices would yield *patterns* supporting reliability and validity, and this was the case (i.e., much higher correlations on than off the diagonals). Were we testing one or two hypotheses with a sample of this size and obtained a *p*-value near .05 or an effect size of $r = .30$, we would place little weight on such findings. The major way sample size does limit our findings regards categorical diagnosis, because we had too few patients with any single Axis II diagnosis to assess reliability or validity of categorical PD diagnoses. This clearly needs to be addressed in future research, particularly if categorical diagnoses are retained in subsequent editions of the DSM.

A second limitation is that it would have been useful to have other data on subjects, such as another PD interview and questionnaire, as well as other criterion variables, so we could compare the validity of different instruments in a single sample. In our prior research, we have found that SWAP-200 diagnoses predict a range of variables, such as family history, developmental history, adaptive functioning, and history of psychiatric hospitalizations, and provide incremental validity above and beyond Axis I diagnosis (see Westen & Harnden-Fischer, 2001; Westen & Shedler, 2000). Such findings, however, now need to be replicated with a larger interview sample using multiple informants.

An additional concern, not specific to this study but more broadly applicable to the method of diagnosis tested here, is that the interview procedure is time-consuming, requiring 2 to 3 hours to complete the interview and an-

TABLE 5. Correlations between SWAP–200 Personality Disorder Dimensional Diagnoses made by Treating Clinicians Adaptive Functioning Ratings made by Independent Interview (N = 15)

	Interview ratings		
Treating Clinician SWAP–200 Diagnosis	**Global assessment of functioning**	**Relationship functioning**	**Employment functioning**
Paranoid	–.40	–.62**	–.56*
Schizoid	–.25	–.40	–.61*
Schizotypal	–.53*	–.69**	–.81+
Antisocial	–.40	–.56*	–.47
Borderline	–.71**	–.64**	–.58*
Histrionic	–.36	–.38	–.33
Narcissistic	–.20	–.33	–.21
Avoidant	–.20	–.23	–.54*
Dependent	–.26	–.12	–.49
Obsessive–Compulsive	.20	.21	.09
Psychological Health Index	.75+	.77+	.75+

Note. +Correlation is significant at the .001 level (2–tailed); **Correlation is significant at the .01 level (2–tailed); *Correlation is significant at the .05 level (2–tailed).

other 30 to 45 minutes to complete the Q-sort procedure following the interview. This raises questions of feasibility. Whether such an approach is worth the time is an empirical question that can only be addressed by future studies bearing on its incremental validity. Personality disorder researchers have come to a consensus that the use of a 60- to 90-minute PD interview is worth the investment of time relative to briefer PD self-report instruments. We suspect that the attempt to assess the roughly 80 diagnostic criteria for the DSM-IV PDs at a rate of 60 to 90 seconds each, as required of the major PD structured interviews, is compromising validity, and further research is necessary to assess costs and benefits of alternative procedures.

PRIMARY FINDINGS

Within these limitations, the primary findings of this study are as follows. First, personality pathology can be reliably assessed from the CDI. Two independent judges can achieve high levels of interrater reliability on SWAP-200 PD scores, including those assessing both DSM-IV Axis II diagnoses and empirically derived personality prototypes. Where the classification system provides nonredundant dimensions, as in our empirically derived classification, interviewers can reliably distinguish very specific dimensions of pathology with high fidelity.

Second, interview and clinical diagnoses using the SWAP-200 correlate strongly with one another, with a median $r>.80$. Given that the former reflect a single-session, cross-sectional assessment and the latter reflect data aggregated over weeks or months of treatment, these correlations are very promising. For diagnostic dimensions expected to be independent (using our empirically derived diagnostic system), convergent validity coefficients average .60 higher than the absolute value of discriminant validity coefficients. By way of comparison are data using the best-validated psychometric instrument in personality research, the NEO-PI-R (McCrae & Costa, 1997), which assesses the 5-Factor Model of personality and is based on over 40 years of research. Correlations between self- and observer reports for the NEO-PI-R are typically in the range of $r = .30$ to .50. This, in turn, is substantially larger than the correlations between most PD interviews and LEAD diagnoses.

Third, the data provide support for a method of deriving nonredundant personality diagnoses empirically from large-*N* samples of patients treated in the community. Previous studies using the SWAP-200 have shared a significant limitation, namely that validity data (e.g., etiological variables, GAF scores, and other measures of adaptive functioning) were all supplied by the same clinician who provided SWAP-200 data. In the present study, different clinicians (interview judges who saw the patient on one occasion, and the patient's treating clinician) converged in their diagnostic judgments, and treating–clinician and interviewer Q-sorts predicted independent assessments of adaptive functioning.

Finally, the data presented here raises questions about whether efforts to maximize reliability and validity of PD diagnosis by structuring the questions clinicians and interviewers ask and keeping clinical judgment to a minimum may be counterproductive. In trying to maximize the reliability of

diagnosis in DSM-IV, Widiger and Frances (1985) concluded that if "interrater reliability is to be achieved, the amount of inference required by the diagnostic criteria must be decreased..." (p. 617). As a result, not only have many PDs become more narrowly and behaviorally defined, but many researchers have called for clinicians to abandon their traditional interviewing methods in favor of structured interviews. The data presented here raises the possibility that we may not need to sacrifice clinical judgment for reliability and validity. Indeed, clinical judgment may be the sine qua non of both.

REFERENCES

Block, J. (1978). *The Q-Sort Method in Personality Assessment and psychiatric research.*

Cassidy, J., & Mohr, J. J. (2002). Unsolvable fear, trauma, and psychopathology: Theory, research, and clinical considerations related to disorganized attachment across the life span. *Clinical Psychology: Science & Practice, 8*(3), 275-298.

Clark, L. A., Livesley, W. J., & Morey, L. (1997). Personality disorder assessment: The challenge of construct validity. *Journal of Personality Disorders, 11*, 205-231.

Dutra, L., Campbell, L., & Westen, D. (in press). Quantifying clinical judgment in the assessment of adolescent psychopathology: Reliability, validity, and factor structure of the Child Behavior Checklist for Clinician-Report.

First, M. B., Spitzer, R. L., Gibbon, M., Williams, J. B. W., Davies, M., Borus, J., Howes, M. J., Kane, J., Pope, H. G., & Rounsaville, B. (1995). The structured clinical interview for DSM-III-R (SCID). Part II: Multi-site test-retest reliability study. *Journal of Personality Disorders, 9*, 92-104.

Gunderson, J. G., Kolb, J. E., & Austin, V. (1981). The Diagnostic Interview for Borderline Patients. *American Journal of Psychiatry, 138*, 896-903.

Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery, and Psychiatry, 23*, 56-62.

Hills, H. A. (1995). Diagnosing personality disorders: An examination of the MMPI-2 and MCMI-II. *Journal of Personality Assessment, 65*, 21-37.

Loranger, A., Susman, V., Oldham, J., & Russakoff, M. (1998). *Personality Disorders Examination (PDE) manual.* Yonkers, NY: DV Communications.

Main, M., Kaplan, N., & Cassidy, J. (1985). Security in infancy, childhood, and adulthood: A move to the level of representation. In I. Bretherton & E. Waters (Eds.), *Growing points of attachment theory and research* (1-2 ed., Vol. 50, pp. 67-104).

McCrae, R., & Costa, P. L. (1997). Personality trait structure as a human universal. *American Psychologist, 52*, 509-516.

Nakash-Eisikovits, O., Dierberger, A., & Westen, D. (2002). A multidimensional meta-analysis of pharmacotherapy for bulimia nervosa: Summarizing the range of outcomes in controlled clinical trials. *Harvard Review of Psychiatry, 10*, 193-211.

Perry, C. J. (1992). Problems and considerations in the valid assessment of personality disorders. *American Journal of Psychiatry, 149*, 1645-1653.

Pfohl, B., & Blum, N. (1995). Obsessive-compulsive personality disorder. In W. J. Livesley (Ed.), *The DSM-IV Personality Disorders.* New York: Guilford.

Pfohl, B., Stangl, D., Zimmerman, M., Bowers, W., & Corenthal. (1995). A structured interview for the DSM-III personality disorders: A preliminary report. *Archives of General Psychiatry, 42*, 591-596.

Pilkonis, P. A., Heape, C. L., Proietti, J. M., Clark, S. W., McDavid, J. D., & Pitts, T. E. (1995). The reliability and validity of two structured diagnostic interviews for personality disorders. *Archives of General Psychiatry, 52*, 1025-1033.

Pilkonis, P. A., Heape, C. L., Ruddy, J., & Serrao, P. (1991). Validity in the diagnosis of personality disorders: The use

of the LEAD standard. *Psychological Assessment, 31*(1), 46-54.

Shedler, J., & Westen, D. (2002). Dimensions of personality pathology: An alternative to the Five Factor Model. *Unpublished manuscript, Emory University.*

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

Spitzer, R. L. (1983). Psychiatric diagnosis: Are clinicians still necessary? *Comprehensive Psychiatry, 24*(5), 399-411.

Westen, D. (1997). Divergences between Axis II instruments and clinical diagnostic procedures: Implications for research and the evolution of Axis II. *American Journal of Psychiatry, 154*, 895-903.

Westen, D. (2002). Clinical diagnostic interview manual. Available at http://www.psychsystems.net/lab.

Westen, D., & Arkowitz-Westen, L. (1998). Limitations of Axis II in diagnosing personality pathology in clinical practice. *American Journal of Psychiatry, 155*, 1767-1771.

Westen, D., & Chang, C. (2000). Adolescent personality pathology: A review. *Adolescent Psychiatry, 25*, 61-100.

Westen, D., & Harnden-Fischer, J. (2001). Classifying eating disorders by personality profiles: Bridging the chasm between Axis I and Axis II. *American Journal of Psychiatry, 158,* 547-562.

Westen, D., Muderrisoglu, S., Fowler, C., Shedler, J., & Koren, D. (1997). Affect regulation and affective experience: Individual differences, group differences, and measurement using a Q-sort procedure. *Journal of Consulting and Clinical Psychology, 65*, 429-439.

Westen, D., & Shedler, J. (1999a). Revising and assessing Axis II, Part 1: Developing a clinically and empirically valid assessment method. *American Journal of Psychiatry, 156*, 258-272.

Westen, D., & Shedler, J. (1999b). Revising and assessing Axis II, Part 2: Toward an empirically based and clinically useful classification of personality disorders. *American Journal of Psychiatry, 156*, 273-285.

Westen, D., & Shedler, J. (2000). A prototype matching approach to personality disorders: Toward DSM-V. *Journal of Personality Disorders, 14*, 109-126.

Westen, D., Shedler, J., Durrett, C., Glass, S., & Martens, A. (2003). Personality diagnoses in adolescence: DSM-IV Axis II diagnosis and an empirically derived alternative. *American Journal of Psychiatry, 160,* 952-966.

Widiger, T., & Frances, A. (1985). The DSM-III personality disorders: Perspectives from psychology. *Archives of General Psychiatry, 42*, 615-623.

Wilberg, T., Dammen, T., & Friis, S. (2000). Comparing personality diagnostic questionnaire-4+ with longitudinal, expert, all data (LEAD) standard diagnoses in a sample with a high prevalence of Axis I and Axis II disorders. *Comprehensive Psychiatry, 41*, 295-302.

Zanarini, M., Frankenberg, F., Chauncey, D., & Gunderson, J. (1987). The Diagnostic Interview of Personality Disorders: Interrater and test-retest reliability. *Comprehensive Psychiatry, 28*, 467-480.

ERRATUM

In an article entitled "Zanarini Rating Scale for Borderline Personality Disorder (ZAN–BPD): A Continuous Measure of DSM–IV Borderline Psychopathology," published in the June issue of *Journal of Personality Disorders* (Vol. 17, No. 3, pp. 233–242), the names of the co–authors were not listed on the article-opening page. The complete listing of author names are as follows:

Mary C. Zanarini, Ed.D., A. Anna Vujanovic, A.B., Elizabeth A. Parachini, B.A., Jennifer L. Boulanger, B.S., Frances R. Frankenburg, M.D., and John Hennen, Ph.D. All authors are from the Laboratory for the Study for Adult Development, McLean Hospital, and the Department of Psychiatry, Harvard Medical School.