ASSESSMENT
1998, Volume 5, Number 4, 333-353

REFINING THE MEASUREMENT OF AXIS II: A Q-SORT PROCEDURE FOR ASSESSING PERSONALITY PATHOLOGY

Jonathan Shedler
Aspen, Colorado

Drew Westen
Harvard Medical School and
The Cambridge Hospital/Cambridge Health Alliance

The measurement of personality disorders (PDs) has proven to be a difficult enterprise. This article describes two initial studies of the validity and reliability of the Shedler-Westen Assessment Procedure (SWAP), a Q-sort procedure that quantifies clinical judgment, which may be useful both for assessing personality pathology and for empirically refining Axis II categories and diagnostic criteria. In the first study, 153 clinicians from a random national sample used a version of the Q-sort to describe either a prototype or actual patient with either a borderline, antisocial, histrionic, or narcissistic personality disorder. Correlations between aggregated prototype and actual patient profiles provided evidence for convergent and discriminant validity, and a cluster-analytic procedure (Q-factor analysis) produced revised criteria for the four disorders that minimized the problem of comorbidity. In Study 2, a pilot sample of patients were interviewed using a clinical research interview that mirrors the way clinicians assess personality and PDs. The study yielded promising results with respect to the possibility of obtaining reliable Q-sort descriptions based on an interview that resembles a clinical interview rather than the direct-question format used in current Axis II structured interviews. It also produced strong correlations between Q-sort descriptions made by interview and those made independently by the treating clinician, further supporting the validity of the instrument. The findings suggest the potential utility of the SWAP as a measure of PDs and as a method for empirically refining Axis II categories and criteria.

Keywords: Axis II, personality disorders, personality assessment, Q-sort, borderline, antisocial, histrionic, narcissistic, prototype

Since the inception of Axis II in the third edition of the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM-III*; American Psychiatric Association, 1980) researchers have made considerable strides in refining personality disorder (PD) diagnoses and in generating a body of empirical literature regarding issues ranging from etiology to prognosis (see Livesley, 1995). A major impediment to progress in their area, however, has been the problematic nature of the instruments

The authors thank the 153 clinicians who provided the data for this study, whose efforts were indispensable.

The authors thank Sandy deSantis for her help in managing the data collection, Howard Shaffer for his comments on an earlier draft, and the over 950 clinicians who helped us refine the measure, including the 153 who took part in the present study.

Correspondence concerning this article and requests for offprints should be addressed to Drew Westen, PhD, Department of Psychiatry, The Cambridge Hospital, 1493 Cambridge Street, Cambridge, MA 02139. E-mail dw@wjh.harvard.edu or Jonathan Shedler, PhD, 42601 E. Highway 82, Aspen, CO 81611. E-mail shedler@ruf.net

designed to assess PDs. Self-report measures have demonstrated only limited validity, correlating minimally with external criteria; interview measures fare better but do not produce validity coefficients nearing those considered acceptable by conventional standards in personality research (see Perry, 1992). For example, Skodol, Oldham, Rosnick, Kellman, and Hyler (1991) found only weak associations between diagnoses made by the Structured Clinical Interview for *DSM-III-R* Personality Disorders (SCID-II; First, Spitzer, Gibbon, & Williams, 1995; Spitzer, Williams, & Gibbon, 1987), the Personality Disorders Examination (Loranger, 1988), and the LEAD (longitudinal expert evaluation using all available data) standard (Spitzer, 1983). Given that these instruments ask similar questions, poor convergent validity is cause for concern. Interrater reliability for current interview measures tends to be high (Zimmerman, 1994), but test-retest reliability is generally fair to poor, particularly if the retest interval is greater than 2 weeks (see First, Spitzer, Gibbon, Williams, Davies et al., 1995; Zimmerman, 1994).

Clinical Assessment and Research Assessment Procedures

These problems doubtless have many causes, but one is of particular import because it may limit the applicability of findings generated by these instruments to clinical work: The measures diverge substantially from the way clinicians actually assess personality disorders. Existing instruments—whether based on self-report questionnaires or structured interviews—share one essential design feature: They attempt to arrive at diagnoses primarily by asking patients direct questions derived from Axis II criteria. This approach had its origins in earlier instruments designed to assess Axis I disorders, such as the Schedule for Affective Disorders and Schizophrenia (SADS; Endicott & Spitzer, 1978) and the Structured Clinical Interview for *DSM-III-R* (SCID; Spitzer, Williams, Gibbon, & First, 1990; Williams et al., 1992), which ask direct questions that are quite similar to those used by clinicians, at least in abbreviated forms. Thus, a patient presenting with depression is asked about suicidality, mood, sleep, vegetative signs, etc. The same kinds of questions may not, however, be as useful for assessing many maladaptive personality patterns. For example, one instrument asks, "Have you ever been told that you seemed like a shallow or superficial kind of person?" to assess histrionic personality disorder (Pfohl, Stangl, Zimmerman, Bowers, & Corenthal, 1985). Another inquires, "Do you feel that your situation is so special that you require preferential treatment?" to assess narcissistic personality disorder (First, Spitzer, Gibbon, & Williams, 1995). The first instruments to assess Axis II disorders, notably the Diagnostic Interview for Borderline Disorders (DIB; Gunderson, Kolb, & Austin, 1981), relied much less on such questions, instead requiring clinical probing and approximately 90 minutes to assess a single disorder. As PD research progressed in the 1980s, however, researchers recognized the importance of providing a more comprehensive diagnostic profile of PD patients, particularly in light of findings of high rates of comorbidity, which led to the development of instruments designed to assess all the PDs in a single interview. An unintended by-product of this methodological development, however, was that a quasi-clinical method gave way to what is largely an interviewer-administered questionnaire method with a few open-ended probes and a request for examples if the patient acknowledges a symptom. What was never tested was the underlying and fundamental assumption that the assessment of Axis I and Axis II disorders can be treated equivalently, and that PD patients can answer direct questions about their pathology accurately.

In two recent studies Westen (1997) found considerable divergence between the way clinicians and research instruments assess personality pathology. The first study surveyed 52 clinical faculty at Harvard Medical School, all with considerable experience in diagnosing and treating personality disorders. The second study was a replication of the first using a national probability sample of over 1,800 experienced psychologists and psychiatrists. In each study, clinicians were asked to rate the extent to which they rely on various methods in making PD diagnoses, among them asking patients direct questions derived from Axis II criteria (as existing research instruments do), listening to patients' narratives about their lives and relationships and making inferences about repetitive patterns, and observing patients' behavior in the consulting room. The clinicians rated each method using a 7-point rating scale (1 = "I rely on it very much" and 7 = "I rely on it very little"). The results indicate that clinicians rely primarily on patients' narrative descriptions of their interactions with significant others ($M = 1.40 \pm 0.87$, $N = 1{,}835$) and on their behavior in the consulting room, particularly their manner of interacting with the interviewer ($M = 1.52 \pm 0.96$). Clinicians find direct questions derived from Axis II criteria of limited use in making Axis II diagnoses ($M = 4.96 \pm 1.82$) but of considerably greater utility in making Axis I diagnoses ($M = 2.67 \pm 1.70$). (All comparisons between methods were significant at $p < .0001$ in both samples). Thus, whereas Axis I instruments mirror clinical procedure, Axis II instruments do not. This pattern of findings emerged regardless of clinicians' theoretical orientation.

Problems With Axis II Translate Into Problems With Research Instruments

Instruments that assess PDs share other design flaws that reflect problems with the *DSM-IV* (American Psychiatric Association, 1994) itself. For example, many characteristics that are clearly continuous in nature (e.g., fears of abandonment, unstable relationships, identity disturbance, impulsivity, affective instability, intense anger, feelings of emptiness, and paranoid or dissociative symptoms–8 of the 9 Borderline Personality Disorder, BPD, criteria) must be coded as categorical (present/absent). Cutoffs for the presence or absence of PD diagnoses are arbitrary, and a slight change in the way a patient answers a question about these criteria can lead to a different diagnosis.

Further, like Axis II, current instruments fail to include many enduring personality problems that bring people to treatment and require clinical intervention (Westen, 1997; Westen & Arkowitz-Westen, in press). Many of these symptoms and patterns, such as recurring problems with intimacy, work, and self-esteem, are captured by neither Axis I nor Axis II and may not be readily assessed as subclinical variants of current Axis II disorders that could be assessed by dimensional measures of the same constructs. For example, many high-functioning patients have fears of abandonment or other attachment-related problems that require clinical attention, but they are clearly not subclinically borderline.

In a recent study (Westen & Arkowitz-Westen, in press), a national sample of psychiatrists and psychologists provided information on randomly selected patients ($N = 714$) they currently were treating for problematic personality patterns, defined as enduring, maladaptive patterns of thought, feeling, motivation, or behavior. Clinicians were asked to check off whether each patient met criteria for each of the Axis II PDs and relevant Axis I disorders, and whether the patient had other clinically significant personality patterns requiring treatment, using a list of personality problems identified in previous research. Only 39.4% of patients being treated for personality pathology had diagnosable Axis II disorders. This percentage was relatively stable across clinicians' theoretical orientations and did not vary substantially when controlling for Axis I diagnosis.

The aim of this paper is to describe initial data on the reliability, validity, and potential utility of a new Q-sort instrument, the Shedler-Westen Assessment Procedure (SWAP), which was developed to address these difficulties. The SWAP was designed to assess personality pathology in a way that more closely resembles the processes clinicians use to diagnose personality pathology and formulate cases, allowing them to target specific processes for treatment. The instrument is intended as both an assessment tool and as an instrument that may be helpful in empirically refining the categories and criteria included on Axis II.

Development of the SWAP

The Q-sort is one of the most successful methodologies that has been employed in the study of normal personality (Block, 1971, 1978; Block, Gjerde, & Block, 1991; Colvin, Block, & Funder, 1995; John & Robins, 1994; Shedler & Block, 1990). Despite its demonstrated value, the Q-sort method has rarely been extended to the study of

PDs. A Q-sort (in the context of personality assessment) is a set of statements that describe personality and psychological functioning. Each statement may describe a given patient well, somewhat, or not at all. Each statement is printed on a separate index card. A clinician or interviewer with thorough knowledge of the patient sorts (rank-orders) the cards into a series of piles based on the degree to which the statements describe the patient, from those that are inapplicable or not descriptive to those that are highly descriptive. This use of the Q-sort relies on the judgments of a clinician-observer rather than on the self-reports of the patient.

In the latest implementation of the SWAP, the SWAP-200, clinicians sort 200 descriptive statements into 8 piles or categories. (The SWAP-167, used in Study 1, had 167 statements sorted into 9 piles.) The first category, which is assigned a value of "0" for data analytic purposes, contains statements that the clinician judges irrelevant or inapplicable to the patient. The last category, which is assigned a value of "7", contains statements that are highly descriptive of the patient; intermediate categories contains statements that apply to varying degrees. In essence, the SWAP-200 provides a numerical score ranging from 0 to 7 for each of 200 personality-descriptive items or statements. The statements provide a standard vocabulary for clinicians to use in expressing their observations and inferences. The distribution of Q-sort items into piles is fixed (i.e., the clinician must assign a specified number of statements to each category), a property of the method that has psychometric advantages discussed in detail by Block (1978).

The use of a standard vocabulary allows clinicians to express observations and inferences in a form that can be (a) quantified, (b) statistically analyzed, and (c) compared with those of other clinician-observers. SWAP statements are written in a manner close to the data (e.g., "tends to be passive and unassertive" or "living arrangements are chaotic and unstable"), and items that require inference about internal processes are stated in clear and unambiguous language (e.g., "is unable to describe important others in a way that conveys a sense of who they are as people; descriptions lack fullness and color," or "tends to blame others for own failures or shortcomings; tends to believe his/her problems are caused by external factors"). Writing items in this way minimizes idiosyncratic and unreliable interpretive leaps. This is similar to the efforts of the Axis II work groups, whose diagnostic criteria have become progressively closer to the data (see Livesley, 1995). The major differences are that we have attempted to operationalize subtle psychological constructs that have typically eluded reliable measurement and have expanded the range of items to capture aspects of functioning of potential clinical importance (such as areas of healthy or adaptive functioning) that Axis II does not address (see also Clark, Livesley, Schroeder, & Irish, 1996).

Development of the SWAP-200 Item Set

The value of a Q-sort depends entirely on the statements that comprise it. An initial item set was drafted by the first author and was revised and refined by both authors over a period of 7 years. The final item set incorporates constructs from a mixture of sources: *DSM-III-R* (American Psychiatric Association, 1987) and *DSM-IV* Axis II criteria; selected Axis I items that could reflect personality disturbance (such as depression and anxiety); clinical literature on PDs written over the past 50 years (e.g., Kernberg, 1984; Kohut, 1971); research on coping and defense mechanisms (Perry & Cooper, 1987; Shedler, Mayman, & Manis, 1993; Vaillant, 1992); research on interpersonal pathology in PD patients (Westen, 1991; Westen, Lohr, Silk, Gold, & Kerber, 1990); research on normal personality traits and psychological health (Block, 1978; McCrae & Costa, 1990); research on the psychological characteristics of PDs conducted since the development of Axis II (see Livesley, 1995); 3-hour pilot interviews in which observers watched videotaped interviews of patients with a range of personality disturbances and tried to describe them using the Q-sort procedure; and clinical observation.

To refine the item set and determine the desired distribution for the Q-sort, the investigators first asked 12 experienced clinicians to rate four patients each using the SWAP items (on a scale from 1 to 9, 1 = "does not describe the patient" and 9 = "is absolutely defining of the patient").

The clinicians were asked to comment on any items that were ambiguous or could not easily be rated for any given patient, to comment on items that were redundant or poorly worded, and to suggest any items that were needed to capture the pathology of the patient but were missing from the item set. All items were then correlated with each other, and items that showed minimal variance or correlated above .80 with any other item were reworded or eliminated. This process was used to create the first generation of the SWAP, the SWAP-167, which included 167 items and was used for the first study reported in this article.

The data from Study 1, described below, were subsequently used to modify the instrument for its next iteration. Once again, items that proved redundant or showed minimal variance in this study were also reworded or eliminated. Perhaps most importantly, the 153 clinicians who participated in the study were asked to comment on the items in the SWAP-167, and their comments were used to modify the instrument. These and other changes led to the development of the 200-item set. Finally, we modified the distribution (for example, switching from 9 piles to 8 by combining the two biggest piles in the distribution so that raters would not have to make fine discriminations between items that are essentially irrelevant to the diagnosis).[1] With these changes, the procedure now takes 45 to 60 minutes following either an interview or based on clinical knowledge of the patient.

Applications of Q-Sort Methodology[2]

Composite descriptions

A major benefit of the Q-sort technique is that it allows personality descriptions provided by different clinicians to be combined to arrive at a single composite personality description for a particular type of patient. This is accomplished by averaging the values assigned to each Q-sort item from multiple clinicians. For example, if a number of experienced clinicians are asked to provide a Q-sort description of a hypothetical, prototypical patient with histrionic PD, their responses can be averaged to obtain a composite description of the prototypical histrionic patient. One fortunate statistical consequence of averaging is that only items ranked highly by all clinicians will have a high ranking in the composite Q-sort. When there is not clinical consensus about an item, the item will not achieve a high ranking in the composite. Thus, by listing the highest-ranking items from the composite of hypothetical, prototypical histrionic patients, one obtains a listing of the psychological features that virtually all clinicians consider important to the diagnosis.

[1] The distribution for the SWAP is a slightly flattened right tail of a normal curve, with a predominance of items placed in Pile 0 (that is, items that are not true of the patient). We chose the distribution on empirical grounds, deriving it from the natural distribution that emerged when averaging the distributions produced spontaneously by raters initially using the measure in rating-scale format. This was also roughly the distribution we anticipated, since most symptoms are untrue of most people; psychopathology is *abnormal.* Most people will not have most symptoms, and a symmetrical distribution will not be appropriate. This would be equally true of most Axis I symptoms if their distributions were plotted, since most people do not have most symptoms. In a college student sample, for example, the distribution of Beck Depression Inventory scores would not be bell-shaped because the average student has a BDI near zero. Further, because two symptoms a patient *does not* have are equally absent, judging which is *more absent* is a question that has no meaning. Consider the task of a judge asked to answer reliably the question of whether a tendency to have idiosyncratic perceptual experiences or a tendency to hoard is *more untrue* of a patient with an adjustment disorder. Both are equally untrue, and forcing the data into a normal distribution would be inappropriate. A normal distribution would be more appropriate if most symptoms had an opposite (e.g., mistrust vs. trust) and if the opposite were as defining of a healthier person as the symptom is of pathology.

[2] This section focuses on empirical applications of the Q-sort method. The reader should be aware that the SWAP-200 may have potential clinical applications as well. For example, both authors have used the SWAP-200 in the context of clinical supervision. The nature of the item set requires the supervisee to reflect on, and render judgments about, a comprehensive range of psychological constructs, and to consider aspects of personality pathology that might otherwise have escaped attention. The item set requires the trainee (and supervisor) to offer psychological descriptions in precise and unambiguous language, without masking areas of confusion behind arcane theoretical terms. The SWAP descriptions provide a rigorous observational database for drawing inferences about personality dynamics, and articulating treatment strategies. The instrument may also prove useful in forensic settings, since instead of administering self-report questionnaires to individuals who know they are being evaluated for competence to stand trial or parental competence, the SWAP-200 allows a quantified, reliable clinical judgment about the patient's personality, which can be compared to national norms of relevant groups, such as recidivists, batterers, pedophiles, etc.

Similarly, if one creates a composite description of a group of actual histrionic patients, only items ranked highly for all patients will have a high ranking in the composite description. Thus, an examination of the highest-ranking items from the composite Q-sort will reveal the important psychological features that these actual histrionic patients have in common. This represents a purely empirical procedure for identifying the psychological features (diagnostic criteria) that characterize histrionic patients.

Similarity of Personality Descriptions Making Up a Composite

Coefficient alpha (Cronbach, 1951) provides a measure of the similarity or agreement (internal consistency) among the multiple Q-sort descriptions that make up the composite. The degree of similarity or agreement between any two Q-sorts is measured by the familiar correlation coefficient. A SWAP-200 Q-sort is one column by 200 rows of data. By correlating two columns of Q-sort data, one obtains a correlation coefficient indicating the similarity or agreement between the two Q-sorts. When more than two columns (more than two correlations) are involved, similarity or agreement is measured by coefficient alpha, a measure that reflects the correlations between all possible pairs of Q-sorts. Thus it becomes possible to ask, for example, to what extent clinicians who describe hypothetical, prototypical patients with borderline PD agree or disagree about the features that constitute the diagnosis. This is an important question, given the debate about whether borderline PD is a distinct and meaningful diagnosis at all. If coefficient alpha is low, this would indicate that different clinicians use the diagnostic term differently, with poor agreement about what constitutes borderline PD. If coefficient alpha is high, this would indicate that there is good agreement among clinicians about the features that make up the diagnosis.

With composite Q-sort profiles based on *actual* patients, coefficient alpha indicates the degree of similarity between these patients. For example, a high alpha coefficient for a composite Q-sort based on actual patients diagnosed with borderline PD would indicate that the patients have important psychological features in common (and therefore represent a coherent diagnostic group). Conversely, a low alpha coefficient would suggest that the patients do not share important psychological characteristics (and that borderline PD is therefore a "catch-all" diagnostic category, as some have claimed, made up of patients who have little in common).

Using the SWAP to Make Diagnoses

The Q-sort method can allow clinicians and researchers to make psychometrically rigorous but clinically sensible diagnoses without over-relying on a direct-question interview format. One way to accomplish this is to ask an appropriate sample of highly experienced clinicians to use the SWAP-200 to describe a hypothetical, prototypical patient who illustrates the personality disorder of interest. The composite prototype description aggregated across a sample of clinicians may then serve as a template for the PD, and Q-sort descriptions of actual patients can be compared to this template to gauge the degree of match. The degree of match or overlap is assessed by a simple correlation coefficient. This method has been extensively used in personality research (see Block, 1978) but has rarely been extended to PDs.

Thus, instead of asking clinicians or research assistants to make diagnoses, which they have difficulty doing reliably, the Q-sort procedure requires them to make *behavioral observations* during a clinical evaluation or over the course of many sessions, and a correlation coefficient assesses the match between the patient's Q-sort profile and the prototype. By collecting a set of prototypes for all the Axis II PDs (or other relevant groups, such as patients who responded or did not respond to serotonin reuptake inhibitors or cognitive-behavioral treatment for depression; batterers with a history of recidivism, etc.), this procedure can yield an MMPI-like profile for a given patient for each PD or other relevant categories (Westen & Shedler, in press-a,b). Like the Minnesota Multiphasic Personality Inventory (MMPI-2; Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989) this is an empirical criterion keying approach, but one based on the observations of skilled observers rather than self-reports.

The Q-sort procedure also yields a narrative description of the patient's most salient diagnostic features, by rank-ordering the items in descending order of diagnostic fit. This is one of the advantages of a Q-sort method, since it not only provides a quantitative profile but also describes the patient in narrative form (using the items from the Q-sort) in standardized language. Thus, the Q-sort provides a quantified clinical case formulation (see Westen, 1998).

Study 1

The goal of the first study was to examine convergent and discriminant validity, by examining whether actual patients with a given diagnosis as described by their treating clinician resemble the prototype for that diagnosis (as described by an independent sample of clinicians) more than they resemble the prototype for other diagnoses. We selected four PDs to study (antisocial, borderline, histrionic, and narcissistic), each from the Cluster B of Axis II, to provide a rigorous test of validity of the instrument.

Participants

We contacted a random sample of 1,000 members of the Clinical Psychology division of the American Psychological Association and asked if they would be willing to participate in a study testing a new instrument for assessing personality disorders. In the initial letter, we asked them whether they were currently treating patients who met criteria for any of the four Cluster B PDs. We also asked their theoretical orientation, years experience post-licensure, number of hours of patients seen per week, primary clinical practice setting, and the socioeconomic status of the patients they treat.

Of the 1,000 contacted, 52% responded, of whom 302 met the following criteria for inclusion in the study: current licensure, more than 2 years' experience post-licensure, and a minimum of 10 hours per week current clinical work. Of those who were invited to participate in the study, 153 provided Q-sort data.[3] The clinicians were an experienced group, with a mean of 15.19 years of experience post licensure (*SD* = 8.0 years, range = 2 to 38 years). Theoretical orientations were cognitive-behavioral (40), psychodynamic (40), family systems (4), eclectic (59), and other (10). Of the 59 who reported an eclectic orientation, 23 identified themselves as primarily cognitive-behavioral, 22 as psychodynamic, 5 as systemic, and 3 as other.

Procedures

Half the clinicians were instructed to provide a SWAP-167 description of a patient they were currently treating who they were certain met the criteria for one of the four PDs of interest. (The diagnosis was pre-selected by the investigators, based on clinicians' prior survey responses, to ensure that the clinician was currently treating such patients.) The other half were asked to use the Q-sort to describe their mental prototype of a patient with one of the four PDs (again pre-selected). The instruction to this second group was as follows (using antisocial here as an example):

> We are asking you to use the SWAP-167 to describe a hypothetical patient with an antisocial personality disorder. We do not want you to describe a real patient. Rather, we are interested in learning what the term "antisocial personality disorder" connotes for you. We would like you to describe a prototypical antisocial patient, a hypothetical person who illustrates antisocial personality disorder in its purest form.

Clinicians were asked to use a 1 through 9 distribution to sort the statements, placing items in Pile 1 that are clearly not true or irrelevant to a description of the patient, and placing items in Pile 9 that are defining of the patient's personality. At the end of the procedure, clinicians were asked to comment on the SWAP-167, and in particular to provide detailed comments about any items they had difficulty applying to their patient, ambiguities in any of the items, and any important personality

[3] The 50% response rate reflects the fact that the procedure takes 45 to 60 minutes to complete, and participants were volunteering their time, so that many were likely daunted when the materials arrived. However, clinicians of the major theoretical orientations participated in roughly equal numbers, and no obvious biases in sampling appeared that could affect the results. A subsequent study in which we were able to offer a small honorarium produced a response rate above 70% for those who returned the survey.

characteristics that they could not describe using the SWAP-167.

Results and Discussion

The study yielded eight aggregate or composite Q-sorts: a composite description of the prototypical patient for each of the four personality disorders (borderline, antisocial, histrionic, and narcissistic), and a composite description of actual patients believed to have one of the four disorders. Approximately 20 clinicians contributed Q-sort descriptions to each of the eight composites.

Regarding the internal consistency of the responses within each category,· the data showed that clinicians clearly share an understanding of each of the PDs as assessed by the internal consistency of the SWAP-200 prototype profiles, and vagaries of diagnosis did not prevent the construction of relatively accurate aggregate descriptions of actual patients. Coefficient alpha was uniformly high, with slightly fewer than 20 judges per prototypical patient composite: antisocial, .97; borderline .95; histrionic, .90; and narcissistic, .90. Alphas remained high but were, predictably, slightly lower for actual patient composite descriptions: .93, .84, .82, and .81, respectively. This indicates that the patients identified as having a given disorder do, in fact, share important features in common, and confirms that their diagnoses were largely accurate. These high alphas give us confidence that the composite descriptions have minimal error, since alpha essentially offers an estimate of true variance and increases with the number of raters because across raters random errors cancel out. Another way to describe these data is to report the average correlation between each patient's profile within each category and the mean profile for that category (subtracting out, of course, the contribution of the patient to the composite). These correlations were again quite high, ranging from r = .45 (narcissistic actual) to r = .79 (antisocial prototype).

Validity

Composite PD descriptions of both prototype and actual patients varied only minimally by the theoretical orientation of the clinician who provided them, although minor differences emerged in some of the composite profiles. For example, with respect to the borderline PD prototype, psychodynamic clinicians were more likely than clinicians in the sample as a whole to give a higher ranking to the items "Tends to feel empty and bored" (composite rank 7.8 vs. 6.6), "Emotions tend to change rapidly and unpredictably from moment to moment" (7.7 vs. 6.8), and "Manages to elicit in others feelings similar to those he or she is experiencing (e.g., when angry, acts in such a way as to provoke anger in others; when anxious, acts in such as way as to induce anxiety in others" (7.2 vs. 5.6).

Evidence for convergent and discriminant validity is provided in Table 1. As can be seen from the table, the numbers along the diagonal, representing the correlation between composite prototypes and aggregated actual patient profiles for each disorder, are high, ranging from .68 to .91. Thus, convergent validity for the SWAP-167 tends to be very strong. For a measure to be valid, however, correlations off the diagonal should be lower in

Table 1
Correlations Among Composite Actual Patient and Prototypical Personality Disorder (PD) Q-sort Descriptions Using the SWAP-167

	Prototype PDs			
Actual Patient PDs	Antisocial	Borderline	Histrionic	Narcissistic
Antisocial	.91	.45	.31	.63
Borderline	.23	.68	.34	.42
Histrionic	.25	.54	.79	.48
Narcissistic	.58	.54	.60	.77

every case than those along the diagonal, which they are. In fact, the average difference between correlations on and off the diagonal is .33, which is impressive given the likely comorbidity of the disorders "in nature" and the overlapping criterion sets in *DSM-III-R*, which was the basis of diagnosis in this study (which preceded the *DSM-IV* by several months).

Discriminant validity was weakest for items comprising the narcissistic profile, suggesting the need for item refinement (see Discussion). Another problematic area was the comorbidity of borderline and histrionic PD, which have shown high estimates of comorbidity in all previous studies as well (see Gunderson, Zanarini, & Kisiel, 1995; Pfohl, 1995). As can be seen from the table, actual borderline patients do not appear particularly histrionic, but actual histrionic patients correlate strongly (r = .54) with the borderline prototype. This suggests that the histrionic construct may itself be indistinct from the borderline construct, at least as it is currently constituted.

Refining Cluster B Categories and Criteria

We thus attempted to use the SWAP-167 to try to refine the Cluster B categories and criteria empirically, to minimize comorbidity that may be an artifact of constructs that are not optimally constituted. To do so, we used a clustering procedure, Q-factor analysis (inverted factor analysis), to see how patients would naturally cluster based on the similarity of their profiles across all 167 items, irrespective of clinician diagnosis. Instead of examining patterns of covariation among statements and grouping together those statements that appear to be variations on a core construct (factor analysis), Q-factor analysis inverts the matrix and treats *cases* (in this case, patients) as the "items" to be factored, and hence groups patients together based on their similarity across all 167 items.

We first entered all patients' profiles into a principal components analysis, initially specifying eigenvalues equal to or greater than 1.[4] The scree plot showed a clear break after 4 factors, where variance explained dropped from 4.15% to 1.97%. We then ran two additional principal components analyses, one specifying a four-factor solution and the other a five-factor solution (so that "noise" would not be forced into the fourth factor) with a varimax (orthogonal) rotation. Both procedures yielded similar solutions with four clear and interpretable Q-factors. Here we report data on the 5-factor solution, which produced the most coherent pattern, retaining the first four factors, which cumulatively accounted for 56.0% of the variance. (For Q-sort typological studies using similar methods, see Block, 1978; Caspi, 1998; Colvin, Block, & Funder, 1995; and Shedler & Block, 1990.)

Tables 2 through 5 show the items from the SWAP-167 that best categorize subjects in these four Q-factors, arranged in descending order of descriptiveness. The numbers in the second column represent the factor scores (obtained by multiple regression) for each item, arranged in descending order of magnitude. These scores reflect the extent to which the item is diagnostic of patients in the Q-factor, and are analogous to the factor scores received by each patient in standard factor analysis; with the matrix inverted, items, not patients, receive factor scores. In other words, whereas in standard factor analysis *patients* receive standardized scores reflecting their value on the factor relative to the mean, Q-factor scores assign standardized scores for each *item* reflecting the extent to which it is diagnostic of the Q-factor.[5]

As can be seen from Table 2, the first Q-factor, *antisocial*, was similar in many respects to both the current antisocial diagnosis and classic

[4] One potential concern regarding this analysis is the relative ratio of items to cases, which in this study is 151 to 167 (since in Q-factor analysis, items and cases are inverted). As noted in the discussion, the Q-factor solution at which we arrived has since replicated in a much larger sample, so it does in fact appear to be stable. There is now considerable debate in the factor-analytic literature about the ratio of items to cases necessary for confidence in a factor solution, and studies with smaller ratios have not produced factor solutions that are appreciably empirically different from those produced with larger ratios. Unfortunately, factor-analytic (or in this case, Q-factor-analytic) studies of PDs do not have the luxury of collecting the kinds of large samples available in undergraduate subject pools; an N of 153 is a large N in comparison to most PD studies.

[5] Taking the mean of the items with the highest loadings provides a similar index, but based on three samples collected to date we have found that the factor scores provide a more clinically and conceptually coherent description.

Table 2
Q-Factor 1: Antisocial

Item	Score
Takes advantage of others; is out for number one; does not appear invested in moral values	3.32
Tends to abuse illicit drugs	2.95
Tends to act impulsively, without regard for consequences	2.82
Work life is chaotic or unstable (e.g., working arrangements seem always temporary, transitional, or ill-defined)	2.56
Has had numerous sexual involvements relative to cultural or subcultural norms; tends to be sexually promiscuous	2.51
Tends to abuse alcohol	2.43
Has little empathy; seems unable to understand or respond to others' needs, wishes, and feelings unless they coincide with his or her own	2.33
Living arrangements are chaotic or unstable (e.g., living arrangements seem always temporary, transitional, or ill-defined; may have no permanent address or no telephone)	2.18
Tends to feel empty or bored	2.10
Tends to be angry or hostile (whether consciously or not)	2.09
Interpersonal relationships tend to be unstable, chaotic, and rapidly changing	2.05
Has a limited ability to label own emotions or to distinguish among them	1.73
Is reckless or accident prone; takes needless risks with own physical safety (whether consciously or unconsciously)	1.70
Tends to take impulsive or ill-considered actions in an effort to manage or escape unpleasant feelings (e.g., may engage in promiscuous sexual activity, start an altercation, break off a relationship, use alcohol or drugs, etc.)	1.62
Tends to get into power struggles	1.61
Tends to have conflicts with authority-figures (e.g., feels he or she must submit, defeat, rebel, etc.)	1.59
Tends to blame others for own failures or shortcomings; tends to feel his or her problems are caused entirely by external factors	1.51
Lacks a stable image of who he or she is or would like to become; long-term goals may be unstable and changing	1.46

Table 3
Q-Factor 2: Emotionally Dysregulated

Item	Score
Tends to feel unhappy, depressed, despondent	2.91
Has low self-esteem; tends to see self in an unfavorable light	2.81
Is unable to soothe or comfort self when distressed; requires involvement of another person to help regulate affect	2.30
Is simultaneously needy of, and rejecting toward, others (e.g., craves intimacy and caring, but tends to reject it when offered)	2.30
Tends to feel ashamed or embarrassed	2.24
Emotions tend to spiral out of control, leading to extremes of excitement, anxiety, sadness, rage, etc.	2.07
Tends to fear abandonment by those who are emotionally significant	2.06
Tends to feel inadequate or inferior; tends to feel like a failure	2.02
Feels helpless and at the mercy of external forces; feels own wishes or actions have little effect	1.97
Tends to be anxious	1.79
Tends to be angry or hostile (whether consciously or not)	1.75
Tends to feel misunderstood or mistreated	1.61
Tends to oscillate between undercontrol and overcontrol of needs and impulses (i.e., needs and wishes are expressed impulsively and with little regard for consequences or else needs and wishes are disavowed and permitted virtually no expression)	1.57
Tends to react to criticism with rage or humiliation	1.55
Tends to feel unworthy and undeserving of success or happiness	1.40
Tends be pessimistic; tends to assume things will not work out	1.39
Tends to be overly needy or dependent; requires excessive reassurance or approval	1.37
Tends to have conflicts with authority-figures (e.g., feels he or she must submit, defeat, rebel, etc.)	1.31
Lacks close friendships and relationships	1.31
Tends to find little or no pleasure, satisfaction, or enjoyment in life's activities	1.24
Tends to blame self or feel responsible for bad things that happen	1.21
Lacks a stable image of who he or she is or would like to become; long-term goals may be unstable and changing	1.19
Appears engaged in a futile effort to elicit approval, acceptance, support, etc. from a parent or parent-figure who cannot or will not provide it	1.16

Table 4
Q-Factor 3: Histrionic

Item	Score
Expresses emotion in exaggerated and theatrical ways	2.89
Becomes attached quickly and intensely; develops feelings, expectations, etc. that are not warranted by the history or context of the relationship	2.78
Tends to be overly needy or dependent; requires excessive reassurance or approval	2.63
Seeks to be the center of attention	2.46
Appears preoccupied with winning the attention or admiration of people he or she perceives as important or high-status	2.22
Emotions tend to spiral out of control, leading to extremes of excitement, anxiety, sadness, rage, etc.	2.12
Fantasizes about finding ideal, perfect love	2.07
Tends to fear abandonment by those who are emotionally significant	2.03
Seems to view others primarily as an audience to witness own importance, brilliance, beauty, etc.	1.83
Emotions tend to change rapidly and unpredictably from moment to moment	1.81
Tends to avoid taking initiative or responsibility for own life; seems to want to be cared for or provided for indefinitely (whether this is conscious or not)	1.77
Tends to be emotionally intrusive; tends not to respect others' autonomy, need for privacy, etc	1.76
Tends to create relationships that repeat or reenact problematic aspects of his or her relationship with a parent	1.56
Tends to be sexually possessive or jealous; tends to be preoccupied with concerns about infidelity (whether real or imagined)	1.55
Tends to be energetic and outgoing	1.43
Manages to elicit in others feelings similar to those he or she is experiencing (e.g., when angry, acts in such a way as to provoke anger in others; when anxious, acts in such a way as to induce anxiety in others)	1.41
Is unable to soothe or comfort self when distressed; requires involvement of another person to help regulate affect	1.38
Perceptions seem glib, global, and impressionistic; has difficulty focusing on specific details	1.30
Tends to have unrealistically idealized views of certain others; sees them as "all good," to the exclusion of commonplace human defects	1.30
Tends to distort beliefs, perceptions, memories, etc. to fit his or her desired view of reality	1.28

Table 5:
Q-Factor 4: Narcissistic

Item	Score
Has low self-esteem; tends to see self in an unfavorable light	2.81
Is unable to soothe or comfort self when distressed; requires involvement of another person to help regulate affect	2.30
Tends to be critical of others	3.09
Appears to feel privileged and entitled; expects preferential treatment	2.30
Is articulate; can express self well in words	2.29
Tends to be oppositional, contrary, or quick to disagree	2.22
Tends to feel misunderstood or mistreated	2.12
Tends to react to criticism with rage or humiliation	2.12
Tends to be self-righteous or moralistic	2.11
Tends to get into power struggles	2.05
Tends to have conflicts with authority-figures (e.g., feels he or she must submit, defeat, rebel, etc.)	1.99
Tends to be conscientious and responsible	1.93
Tends to blame others for own failures or shortcomings; tends to feel his or her problems are caused entirely by external factors	1.69
Manages to elicit in others feelings similar to those he or she is experiencing (e.g., when angry, acts in such a way as to provoke anger in others; when anxious, acts in such a way as to induce anxiety in others)	1.67
Tends to elicit extreme reactions or stir up strong feelings in others	1.67
Tends to think in abstract and intellectualized terms, even in matters of personal import	1.60
Is able to use his or her talents, abilities, and energy effectively and productively	1.53
Is quick to assume that others wish to harm or take advantage of him or her; tends to perceive malevolent intent in others' words and actions	1.49
Has little empathy; seems unable to understand or respond to others' needs, wishes, and feelings unless they coincide with his or her own	1.49
Is simultaneously needy of, and rejecting toward, others (e.g., craves intimacy and caring, but tends to reject it when offered)	1.41
Tends to become absorbed in details, often to the point that he or she misses what is significant in the situation	1.30
Tends to see self as logical and rational, uninfluenced by emotion; prefers to operate as if emotions were irrelevant or inconsequential	1.28
Is invested in seeing self as psychologically healthy and well-adjusted despite clear evidence of problems	1.24
Tends to have unrealistically devalued views of certain others; sees them as "all bad," to the exclusion of any positive qualities	1.21
Seems to view others primarily as an audience to witness own importance, brilliance, beauty, etc.	1.21

descriptions of psychopathy (Cleckley, 1941; Hare & Hart, 1995) and sociopathy (Robins, 1966). The second, which we labeled *emotionally dysregulated*, included patients with a mix of current diagnoses, but was comprised mostly of a subset of patients currently diagnosed with borderline PD. These patients tend to be acutely and intensely dysphoric. They seek others to help them regulate their poorly modulated affects (Table 3) but do so in ways that tend not to work (such as simultaneously clinging to and rejecting them). What is most striking about the difference between this Q-factor and current Axis II borderline PD criteria is its emphasis on these patients' *pain*. The third Q-factor was a refined *histrionic* category, which included most patients diagnosed by their clinicians with histrionic PD, as well as a subset of other patients, mostly diagnosed with narcissistic PD. As can be seen, these patients are defined not only by many of the histrionic criteria currently in Axis II but also by several current borderline criteria and some narcissistic criteria that have typically made distinguishing between narcissistic and histrionic patients difficult (Table 4). The final Q-factor was a refined *narcissistic* diagnosis (Table 5), which included primarily narcissistic patients along with a small number of patients diagnosed by their treating clinicians as antisocial or borderline. Aside from its clinical coherence, one virtue of this categorization is that the four Q-factors are orthogonal, which suggests that using clustering procedures of this sort, comorbidity may be minimized even while clinical coherence is maintained.

As a first test of the validity of this empirically grounded typology, we examined the correlations between the four Q-factors and the composite prototypes of the current Axis II disorders. In other words, we tested the degree to which these empirically sorted patient groups match clinicians' prototypes of current Axis II categories. As can be seen in Table 6, *even using current DSM prototypes as the criterion validity variables*, the revised categories did substantially better than the *DSM* versions. The antisocial Q-factor correlated with the SWAP-167 antisocial prototype at $r = .84$ but no greater than $r = .39$ (narcissistic) with any other prototype. The emotionally dysregulated Q-factor correlated with the borderline prototype at $r = .61$ but no higher than .06 with any other prototype. As far as we know, this is the only study that has produced a borderline PD group that did not show comorbidity with other disorders. The histrionic Q-factor correlated with the histrionic prototype at $r = .86$ and with the borderline and narcissistic prototypes at $r = .51$ and $r = .54$, as one would expect since it includes many features common to all three disorders as currently conceptualized. The narcissistic Q-factor correlated with the narcissistic prototype at $r = .55$, whereas it correlated only negligibly with the other three prototypes, again showing impressive discriminant validity.

Study 2

The goal of the second study was to assess the reliability of the SWAP-200 by determining whether two clinician-judges using a semi-structured interview that resembles a clinical evaluation could

Table 6
Correlation Between Q-Factors and SWAP-167 Current Axis II Prototypes

	Axis II Prototypes			
Q-factors	Antisocial	Borderline	Histrionic	Narcissistic
Antisocial-psychopathic	**0.83**	0.29	0.29	0.39
Emotionally dysregulated	−0.01	**0.61**	0.06	0.06
Histrionic	0.20	0.51	**0.86**	0.54
Narcissistic	0.27	0.02	0.11	**0.55**

make reliable observations. Establishing interrater reliability with an interview of this sort is much more difficult than with current instruments, since clinicians are drawing inferences about 200 variables based on patients' narratives, rather than determining whether 6 to 10 symptoms are present for each diagnosis based on direct questions, many of which yield yes or no answers from respondents that can inflate estimates of interrater reliability.

Participants

Participants were from The Cambridge Hospital or from the private practice of clinicians involved in the project or willing to refer patients to the study. The sample size (N = 8) is comparable to most pilot studies of interrater reliability in PD research. Patients were recruited by their therapists, who informed them of the study and asked if they would give their permission to be contacted by the researcher. Of the 9 patients contacted, 8 agreed to participate. The sample consisted of 5 women and 3 men, with mean age of 32 years (SD = 7.11 years, range = 19-43 years). Primary diagnoses based on *DSM-IV* (as provided by one of us [D. W.] after watching the interviews and consulting the diagnoses provided by the other judges) included borderline PD (n = 3), adjustment disorder (n = 2), antisocial PD (n = 1), major depressive disorder (n = 1), dysthymia (n = 1), PD NOS (n = 2), and dissociative disorder NOS (n = 2). (Several participants received diagnoses on both Axis I and II.) Patients were paid $25 for their participation.

Procedures

Patients were interviewed by one of the authors or by a clinical psychology fellow, and the interviews were videotaped so that the patient could be evaluated by multiple clinician-judges. Two clinician-judges provided Q-sort descriptions of all 8 patients using the SWAP-200 (which was developed subsequent to the first study) and were blind to all data including diagnosis (and, of course, each other's Q-sort descriptions of the patient). Because participants were outpatients, clinician-judges had no prior clinical interactions with them. Clinician-judges were the second author (who has over 15 years clinical experience) and two psychology fellows who had just completed their predoctoral internship (with approximately 4 years clinical experience each).[6] Training in the SWAP-200 was minimal, involving observation and discussion of three patient interviews. To provide additional validity data, the participants' therapists independently provided Q-sort descriptions of their patients based on their clinical experience with them.

Participants were administered a 2- to 3-hour videotaped interview developed by the second author, called the Personality Diagnostic Interview. The interview was designed to reflect the kind of clinical interview process clinicians actually use to diagnose personality disturbances. The interview proceeds by asking patients to provide narratives about themselves, about what brought them in for treatment, about significant relationships from the past and present, about their work history, about particularly stressful or difficult times in their lives, about their moods and emotions, and about their characteristic ways of thinking. Participants are first asked broad questions, such as, "Can you tell me about your romantic relationships–what have they been like?" Following their responses to these general questions (the main purpose of which is to assess conscious attitudes and to prime specific memories), they are then asked to describe two to three incidents, with instructions such as the following: "Now I'd like you to describe a specific encounter with your husband, something that stands out. It can be an incident that's typical of your relationship, really meaningful, really good, really bad–whatever comes to mind." 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 probes resemble those used by clinicians as well as those used for assessing Thematic Apperception Test responses, core conflictual relationship themes (Luborsky & Crits-Christoph, 1990), and adult attachment patterns (Main, Kaplan, & Cassidy, 1985).

[6] The second author and one fellow were the primary clinician-judges, except where one or the other was the referring therapist, in which case the other fellow was the second clinician-judge.

Unlike current Axis II interviews (but more like clinical practice), when probe questions related to specific Axis II disorders are used, these questions typically *follow* narrative examples rather than precede them. For example, an interviewer who suspects paranoid dynamics after a patient describes two encounters in which he felt like someone was trying to hurt or take advantage of him might ask, "Do you find that people often act that way?" "Do you often feel that way?" or "As we've been talking, have you had any thoughts or worries about what I might do with the information you're giving me?" This is very different from leading with the question, "Are you often suspicious of people's motives?" and following with a request for an example if the patient answers in the affirmative.

Results and Discussion

With respect to interrater reliability, the average correlation between SWAP-200 profiles by clinician-judges who observed the same interview (in person or on videotape) was .61 (Pearson's *r*), which is relatively high for Q-sort data, where multiple judges are typically required in order to achieve acceptable reliability (e.g., Block, 1978). The Spearman-Brown-corrected reliability for the two judges combined was .75. This suggests that two clinician-judges independently describing a patient using the SWAP-200 from an interview can produce reliable results for research purposes if their responses are averaged, although clearly further work would be useful to bring the corrected reliability coefficient to .80.

Study 2 provided further evidence for the validity of the measure as well. The average correlation between the composite interview-based Q-sort descriptions and independent Q-sort descriptions provided by the patients' therapists was .54, which is a relatively strong observer-observer validity coefficient by research standards in personality (e.g., McCrae & Costa, 1990). Given the limits on the reliability of the clinician Q-sorts, this value is also likely to be an underestimate of the actual validity of the instrument. In contrast, correlations of the profiles of patients who did not share a diagnosis (according to the diagnoses listed by the two clinician-judges and the therapist, where available) were quite low (average *r* = .14), suggesting that the interrater reliability and validity estimates obtained did not capitalize on spurious random correlations. The data also suggested that at least some minimal prior training in the procedure is useful, since the only correlations between interview and therapist Q-sort descriptions below *r* = .50 were for two patients whose therapist had no prior training on the instrument.

General Discussion

The data from these two studies provide initial support for the validity and reliability of the SWAP procedure and suggest its potential promise as a method for empirically refining Axis II categories and criteria.

Validity, Reliability, and Potential Utility for Refining Diagnostic Criteria

Using the progenitor of the SWAP-200, the SWAP-167, we found strong evidence for convergent validity, with correlations between aggregated prototype descriptions and aggregated profiles of actual patients for each diagnosis ranging from .68 to .91. With respect to discriminant validity, the data were clear—a .33 discrepancy between correlations on and off the diagonal, which is strong by most standards of convergent-discriminant validity—but less powerful. The major problem with discriminant validity using this initial 167-item sort was with narcissistic PD items. (Interestingly, narcissistic PD was revised more heavily than other PDs in *DSM-IV* for precisely the same reason.) Based on these data, we carefully examined and revised items associated with narcissistic PD to be certain that any convergence between diagnostic categories reflected genuine similarity rather than poor item wording. (For example, in the SWAP-167, we had not adequately distinguished the haughty devaluation of others characteristic of narcissistic patients and the tendency to perceive people at times as all-bad and lacking any positive qualities characteristic of borderline patients.) A subsequent study just completed using the SWAP-200 with a sample of 797 patients with PDs (Westen & Shedler, in press-a) suggested that our efforts in this respect were successful: The correlation between actual and prototypic narcissistic patients was $r = .79$, whereas correlations between actual narcissistic patients and the other Cluster B prototypes ranged from $r = .32$ to $r = .47$ (antisocial). Study 2 provided further preliminary evidence for validity, yielding a .54 correlation between interview-based and therapist-based profiles for individual patients, which is a strong observer-observer correlation, although this will need to be replicated with a larger sample size.

The obtained reliability coefficients are acceptable or nearly acceptable (corrected $r = .75$), but they clearly could be stronger. We have reason to believe they were biased downward by two sources of unreliability not intrinsic to the method. First, we had clinician-judges and therapists consider the past 2 years of the patient's functioning when describing the patient, which we discovered was frequently a source of unreliability because most of the patients had been in continuous treatment for months or years. Q-sort judges were sometimes confused about how to describe a patient who had improved with treatment during the last 2 years, and therapists often realized after watching the interview subsequent to completing their Q-sort description of the patient that they had forgotten how symptomatic the patient had been 2 years earlier, which the interview judges had scored. Thus, in future research, the time frame should be shorter, particularly if the instrument is being used for treatment outcome research. Second, interview judges received minimal training, and some of the therapists received none at all. In fact, in our latest series of 5 patients, the corrected correlation has risen to .81, which also suggests that the relatively small size of the reliability sample reported in Study 2 is unlikely to pose problems for generalizability.

Like recent efforts by Clark et al. (1996) and Morey (1988), Study 1 also attempted to use statistical procedures to refine Axis II diagnoses empirically. The four-factor solution produced by Q-analysis produced four coherent, clinically sensible categories resembling the current Axis II diagnoses, except that they (a) were orthogonal, hence minimizing diagnostic overlap; (b) did not necessitate arbitrary elimination of core items from diagnoses to minimize comorbidity (notably empathy and substance abuse in the antisocial Q-factor); (c) isolated an emotionally dysregulated category that purified the borderline diagnosis and eliminated its comorbidity with all of the other Cluster B disorders; and (d) modified the histrionic diagnosis to include narcissistic and borderline features that have been steadily removed from Axis II over recent *DSM* editions (see Pfohl, 1995) to minimize comorbidity but have produced diagnostic criteria for a non-borderline, non-narcissistic, but nevertheless disturbed version of a "hysterical character" (from which the diagnosis was initially in large measure derived) that may not exist in nature. Our most recent $N = 797$ study using the SWAP-200, which included patients in all 10 *DSM-IV* PD categories and those in the appendix, replicated these findings, it similarly distinguished an emotionally dysregulated Q-factor (comprised primarily of patients currently diagnosed by their treating clinicians as borderline) and a histrionic Q-factor orthogonal to it that had substantial borderline and narcissistic features.

Relative to current research diagnostic procedures, the SWAP has a number of potential advantages: (a) it relies on expert judgment rather than self-report; (b) it uses a criterion keyed, prototype-matching approach to diagnosis; (c) it weights items in terms of their diagnostic fit by virtue of its correlational prototype-matching process, which better matches data showing that some criteria are in fact more diagnostic than others (Davis, Blashfield, & McElroy, 1993); (d) it does not arbitrarily code all diagnostic criteria as dichotomous (present/absent) as in the *DSM-IV* (e.g., how much rejection sensitivity must a patient display to reach the threshold for that criterion for borderline PD?); (e) it relies on interviewing procedures that can be used by competent clinicians; (f) it can be used to diagnosis the entire spectrum of personality pathology, from the problematic but less pathological patterns of thought, feeling, motivation, and behavior that lead most patients with personality problems to enter into treatment (Westen & Arkowitz-Westen, in press) to those that lead to the kind of severe dysfunction seen in the current PDs; (g) its item set includes many subtle psychological processes (such as ways of experiencing the self and others and ways of regulating affects) that are not assessed in current self-reports and structured interviews; and (h) it can provide not only diagnostic labels but a quantified case formulation of a patient's personality functioning in narrative form.

Limitations

The studies reported here have a number of limitations. The first is that sample sizes were relatively small, and hence the findings must be considered in that light. With respect to Study 1, however, as noted above, we have just completed a study with a sample size close to 800 and have achieved even stronger estimates of convergent and discriminant validity across all 10 *DSM-IV* PDs, with the average difference between correlations on and off the diagonal measured two different ways greater than .50 (Westen & Shedler, in press-a).

Second, with respect to the Q-factor analysis, the design was biased in such a way as to be likely to replicate the current Axis II taxonomy, since clinicians were asked to describe patients who fit into the current Cluster B categories. We deliberately chose this strategy for this study so that we could establish the validity of the measure using current categories as the rubber standard (since there is no gold standard). The comorbidity found in this, as in every other sample, made this strategy less problematic, since the current taxonomy does not produce many "pure" types in reality, and the fact that the Q-factor analysis did *not* simply replicate the current taxonomy but instead suggested what seem like conceptually and clinically sensible syndromes points to the potential utility of the method. These syndromes, like the scales similarly assessed through criterion matching using the MMPI-2, can be treated either dimensionally or categorically, by converting participants' scores to *T* scores and selecting cutoffs above which patients are considered to have the disorder (Westen & Shedler, in press-b). Patients with elevated scores below those cutoffs on a scale can be described as having "features" of the disorder, much as clinicians currently describe patients (e.g., borderline PD with antisocial features).

Third, a critic might ask whether the SWAP-167 or SWAP-200 include the right items: Perhaps we obtained the results we did because we used an idiosyncratic instrument. Unfortunately, neither we nor anyone else can demonstrate that we have included all the necessary items and have not included items that are unnecessary. What we *can* say is that (a) we used the methods for scale construction that personality psychologists consider the best methods, such as making successive approximations, trying them out, and eliminating redundant items; (b) we included items from a broad range of sources; (c) we have had, at this point, over 950 clinician-consultants use the instrument and give us feedback about items that are vague or unclear, items that are redundant, or statements that they wished to make about their patient but could not because no item appropriately covered them; (d) the item set makes use of the wisdom of all the individuals who contributed to the last two versions of Axis II criteria but is more comprehensive because it includes items covering all the Axis II criteria plus roughly 130 others; and (e) in our most recent study using the SWAP-200, we asked clinicians to rate the comprehensiveness of the item set and received clear confirmation that we had, in fact, captured the major dimensions on which clinicians assess the personality patterns of their patients. When we asked clinicians to rate the comprehensiveness of the item set, over 75% responded that they were able to describe "most of what is important about the patient's personality" (as opposed to "some of what is most important"–24%–and "little" or "none"–less than 1% for these latter two categories combined).

Fourth, the data were provided by clinicians, who could have biases that limit the reliability of the information they provided. Several considerations limit the impact of this criticism. (a) All observers have biases. Ideally, one would want to rely on as many credible sources as possible, and future research should clearly employ a multi-trait, multimethod approach to assess the validity of the instrument. Nevertheless, we believe the judgments of experts with an average of over 15 years experience who have known a patient over time are certainly to be taken as seriously as either self-reports or judgments made in 30 to 90 minutes by research assistants or interviewers using brief structured interviews that rely on direct questions as the primary mode of gathering information. This is particularly true given the potential confounds of state and trait that make assessment of PDs especially difficult (e.g., Tyrer, 1995; Zimmerman, 1994). One would expect that knowing a patient over time should limit diagnostic noise reflecting the vagaries of current Axis I state conditions that can bias judgments when diagnosis rests exclusively on a brief cross-sectional snapshot of a patient at a single time. (b) The alphas showing the internal consistency of clinicians' descriptions of patients sharing a diagnosis demonstrate that the instrument can in fact be used reliably by clinicians to provide a composite portrait of a disorder. This does not mean that an individual clinician's description of a single patient should be assumed to be reliable; it simply means that if one averages across the descriptions provided by a large enough group of clinicians, one can develop an accurate portrait of a disorder. The high alphas we produced–above .80 in all cases–suggest that we relied upon a large enough sample to accomplish this. That these alphas are so high is in some respects surprising given that clinicians varied substantially in theoretical orientation, training, etc. (c) As documented by survey data (Westen, 1997), the gulf is wide between clinical and research approaches to PDs. If the *DSM* is to guide clinical diagnosis, it should have clinical relevance, and we know of no way better way to guarantee its fidelity to clinical reality than to harness clinical observation to refine it. This kind of collaboration between researchers and practitioners is now a cornerstone of the MacArthur-funded American Psychiatric Association research practice networks.

A related objection is that the validity data may be artifactual, since clinicians describing actual patients may simply have used their implicit prototypes to guide their descriptions. Several factors militate against this criticism as well. (a) The SWAP-167 includes 167 items, whereas any given Axis II diagnosis only includes 8 to 10 criteria. Thus, clinician-respondents in this study were on their own for the other 150 + descriptions. In our most recent study we assessed convergent and discriminant validity a second way, by correlating single-item Likert-type ratings of the extent to which the patient has characteristics of each of the PDs with dimensional SWAP-200 PD scales created by correlating 200 items per patient with prototype profiles generated by a second set of independent judges, and produced identical findings. (b) Clinicians of different theoretical orientations tended to view patients within but not across diagnoses similarly, despite their very disparate theoretical views. (c) Clinicians did not, in fact, simply reproduce the criteria from *DSM-IV* in describing actual patients, suggesting that they were in fact describing their patients when asked to do so and not idealized prototypes. For example, the composite description of borderline patients bore only a family resemblance to the *DSM-IV* description, with much greater emphasis on patients' dysphoria, including despondency, feelings of inadequacy, and anxiety. (d) The Q-factor analysis did not replicate Axis II precisely, and in fact suggested a different categorization of borderline and histrionic PD patients. (e) Although composite descriptions of actual patients correlated highly with composite prototypes of the same disorder, the correlations between any two patients within a given category ranged from .00 to as high as .80. This tremendous variation suggests that clinicians who were instructed to describe an actual patient were not simply describing prototypes and ignoring the attributes of the patient in front of them. (f) In Study 2, Q-sort profiles based on research interviews correlated strongly with clinician Q-sort descriptions of the same patient despite the fact that clinicians were not asked to select a patient with any particular diagnosis and clinicians were blind as to what, if any, Axis II diagnosis the treating clinician believed the patient to have. (g) The problem with relying primarily on reports from a single source (in this case, clinicians) is not specific to this study, since the overwhelming majority of studies of personality and PDs have relied exclusively on two methods–either an interviewer's judgment after a brief clinical interview or on self-reports–neither of which we believe is superior to quantified data obtained from clinicians who have worked with the patient over time. Nevertheless, the next step is clearly a similar cluster-analytic study using a large random sample of patients treated for personality pathology that relies on a mixture of clinician-, self-, and informant-reports.

A final question is whether the validation procedures used here are sufficient for establishing the validity of the instrument. Does the method of establishing convergent and discriminant validity here provide a rigorous enough test of the validity of the instrument? And would similar results have emerged with any other measure of personality disorders? Similar results do not, in fact, emerge with other measures. The answers patients provide to questions constructed to assess borderline PD in self-report measures empirically do not discriminate these patients from patients with almost any other PD diagnosis. One advantage of a method that relies on expert observers is that it does not depend on patients for whom lack of self-knowledge is pathognomonic to describe themselves accurately. Further, in our subsequent study (N = 797) using the SWAP-200, we found even stronger evidence of validity using both the same method used here and other methods that do not rely on a categorical diagnosis by the clinicians. Clearly, however, future research is necessary to assess whether SWAP profiles made by interview can predict relevant variables, such as diagnoses made independently by the treating clinician, SWAP profiles made by the treating clinician, informant data, self-report data on relatively objective behaviors that require less inference on the part of the patient (such as the number of times the patient has socialized with other people in the last week, the number of physical fights the patient has had in the last week, etc.), and behavioral data such as data from beeper studies.

References

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

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

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

Block, J. (1971). *Lives through time.* Berkeley, CA: Bancroft.

Block, J. (1978). *The Q-sort method in personality assessment and psychiatric research.* Palo Alto, CA: Consulting Psychologists Press.

Block, J. H., Gjerde, P., & Block, J. H. (1991). Personality antecedents of depressive tendencies in 18-year-olds: A prospective study. *Journal oJ Personality and Social Psychology, 60,* 726-738.

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *Minnesota Multiphasic Personality Inventory-2 (MMPI-2): Manual for administration and scoring.* Minneapolis: University of Minnesota Press.

Caspi, A. (1998). Personality development across the lifespan. In W. Damon (Ed.), *Handbook of child psychology,* Vol. 3, *Social, emotional, and personality development* (N. Eisenberg, Vol. Ed.) (pp. 311-388). New York: Wiley.

Clark, L. A., Livesley, W. J., Schroeder, M., & Irish, S. (1996). Convergence of two systems for assessing specific traits of personality disorder. *Psychological Assessment, 8,* 294-303.

Cleckley, H. (1941). *The mask of sanity.* St. Louis, MO: Mosby.

Colvin, R., Block, J., & Funder, D. (1995). Overly positive self-evaluations and personality: Negative implications for mental health. *Journal of Personality and Social Psychology, 68,* 1152-1162.

Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297-334.

Davis, R., Blashfield, R., & McElroy, R. (1993). Weighting criteria in the diagnosis of a personality disorder: A demonstration. *Journal of Abnormal Psychology, 102,* 319-322.

Endicott, J., & Spitzer, R. (1978). A diagnostic interview: The Schedule for Affective Disorders and Schizophrenia. *Archives of General Psychiatry, 35,* 837-844.

First, M., Spitzer, R., Gibbon, M., & Williams, J. (1995). The Structured Clinical Interview for *DSM-III-R* Personality Disorders (SCID-II). Part I: Description. *Journal of Personality Disorders, 9,* 83-91.

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

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

Gunderson, J., Zanarini, M., & Kisiel, C. (1995). Borderline personality disorder. In W. J. Livesley (Ed.), *The DSM-IV personality disorders* (pp. 141-157). New York: Guilford.

Hare, R., & Hart, S. (1995). Commentary on antisocial personality disorders: The *DSM-IV* field trial. In W. J. Livesley (Ed.), *The DSM-IV personality disorders* (pp. 127-134). New York: Guilford.

John, O., & Robins, R. (1994). Accuracy and bias in self-perception: Individual differences in self-enhancement and the role of narcissism. *Journal of Personality and Social Psychology, 66,* 206-219.

Kernberg, O. (1984). *Severe personality disorders.* New Haven, CT: Yale University Press.

Kohut, H. (1971). *The analysis of the self.* New York: International Universities Press.

Livesley, W. J. (Ed.). (1995). *The DSM-IV personality disorders.* New York: Guilford.

Loranger, A. (1988). *Personality Disorders Examination (PDE) manual.* Yonkers, NY: DV Communications.

Luborsky, L., & Crits-Christoph, P. (1990). *Understanding transference: The core conflictual relationship theme method.* New York: Basic Books.

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. *Monographs of the Society for Research in Child Development, 50.* (No. 1-2) 67-104.

McCrae, R. R., & Costa, P. T., Jr. (1990). *Personality in adulthood.* New York: Guilford.

Morey, L. (1988). Personality disorders in *DSM-III* and *DSM-III-R*: Convergence, coverage, and internal consistency. *American Journal of Psychiatry, 145,* 573-577.

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

Perry, J. C., & Cooper, S. H. (1987). Empirical studies of psychological defense mechanisms. In R. Michels & J. O. Cavenar, Jr. (Eds.), *Psychiatry.*.Philadelphia: Lippincott.

Pfohl, B. (1995). Histrionic personality disorder. In W. J. Livesley (Ed.), *The DSM-IV personality disorders* (pp. 173-192). New York: Guilford.

Pfohl, B., Stangl, D., Zimmerman, M., Bowers, W., & Corenthal, C. (1985). A structured interview for the *DSM-III* personality disorders: A preliminary report. *Archives of General Psychiatry, 42,* 591-596.

Robins, L. (1966). *Deviant children grown up.* Baltimore: Williams and Wilkins.

Shedler, J., & Block, J. (1990). Adolescent drug use and psychological health: A longitudinal inquiry. *American Psychologist, 45,* 612-630.

Shedler, J., Mayman, M., & Manis, M. (1993). The illusion of mental health. *American Psychologist, 48,* 1117-1131.

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

Spitzer, R., Williams, J., & Gibbon, M. (1987). *Structured Clinical Interview for DSM-III-R Personality Disorders (SCID-II).* New York: New York State Psychiatric Association, Biometrics Research.

Spitzer, R., Williams, J., Gibbon, M., & First, M. (1990). *Structured Clinical Interview for DSM-III-R (SCID).* Washington, DC: American Psychiatric Press.

Tyrer, P. (1995). Are personality disorders well classified in *DSM-IV*? In W. J. Livesley (Ed.), *The DSM-IV personality disorders* (pp. 29-44). New York: Guilford.

Vaillant, G. (Ed.). (1992). *Ego mechanisms of defense: A guide for clinicians and researchers.* Washington, DC: American Psychiatric Press.

Westen, D. (1991). Social cognition and object relations. *Psychological Bulletin, 109,* 429-455.

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. (1998). Case formulation and personality diagnosis: Two processes or one? In J. Barron (Ed.), *Making diagnosis meaningful* (pp. 111-138). Washington, DC: American Psychological Association Press.

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

Westen, D., Lohr, N., Silk, K., Gold, L., & Kerber, K. (1990). Object relations and social cognition in borderlines, major depressives, and normals: A TAT analysis. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 2,* 355-364.

Westen, D., & Shedler, J. (in press-a). Revising and assessing Axis II: I. Developing a clinically and empirically valid method. *American Journal of Psychiatry.*

Westen, D., & Shedler, J. (in press-b). Revising and assessing Axis II: II. Toward an empirically and clinically sensible taxonomy of personality disorders. *American Journal of Psychiatry.*

Williams, J., Gibbon, M., First, M., Spitzer, R., Davies, M., Borus, J., Howes, M., Kane, J., Pope, H., Rounsaville, B., & Wittchen, H. (1992). The Structured Clinical Interview for *DSM-III-R* (SCID): II. Multisite test-retest reliability. *Archives of General Psychiatry, 49,* 630-636.

Zimmerman, M. (1994). Diagnosing personality disorders: A review of issues and research methods. *Archives of General Psychiatry, 51,* 225-245.