Divergences Between Clinical and Research Methods for Assessing Personality Disorders: Implications for Research and the Evolution of Axis II

Drew Westen, Ph.D.

Objective: *The purpose of this study was to examine the extent to which instruments for assessing axis II diverge from clinical diagnostic processes.* *Method:* *Subjects in the first study were 52 clinicians with experience in assessment and treatment of patients with personality disorders, who were surveyed about the methods they use in clinical practice to make diagnoses and other aspects of the diagnostic process. A second study replicated the major findings with a random national sample of 1,901 experienced psychiatrists and psychologists.* *Results:* *Whereas current instruments rely primarily on direct questions derived from DSM-IV, clinicians of every theoretical persuasion found direct questions useful for assessing axis I disorders but only marginally so for axis II. They made axis II diagnoses, instead, by listening to patients describe interpersonal interactions and observing their behavior with the interviewer. In contrast to findings with current research instruments, most patients with personality disorders in clinical practice receive only one axis II diagnosis, and if they receive more than one, one is considered primary. Clinicians reported treating a substantial number of patients for enduring personality patterns that current axis II instruments do not assess, many of which meet neither axis I nor axis II criteria, notably problems with relatedness, work, self-esteem, and chronic subclinical depressive traits.* *Conclusions:* *Measures of axis II were constructed by using a model derived from axis I instruments that diverges from clinical diagnostic procedures in a way that may be problematic for the assessment of personality disorders and the development of a more clinically and empirically sound taxonomy.*

(Am J Psychiatry 1997; 154:895–903)

From the start, axis II has been the subject of considerable controversy (1–4). The task forces charged with revising axis II since DSM-III have paid careful attention to the results of scientific studies in their revisions of diagnostic categories and criteria (DSM-IV). Nevertheless, difficulties in the categories and criterion sets that constitute axis II remain, such as overlap of diagnostic criteria, lack of clear empirical procedures for selecting disorders and criteria, and questions about the syndromal nature of personality disorders. These have led some, including members of the axis II work group, to argue for a dimensional rather than a categorical approach to personality disorder diagnosis (5) or for prioritization of personality disorder diagnoses so that some take precedence over others (6).

PROBLEMS WITH VALIDITY OF INSTRUMENTS FOR ASSESSING PERSONALITY DISORDERS

Developing instruments for the valid assessment of personality disorders has also proven a difficult task. In 1992 Perry (7) reviewed the evidence on existing axis II structured interviews and questionnaires and found that although most demonstrated considerable reliability (test-retest, interrater, or both), none had demonstrated acceptable evidence that it validly assessed the

Received April 24, 1995; revisions received Oct. 4, 1995, and March 22, Sept. 3, and Dec. 16, 1996; accepted Jan. 6, 1997. From the Department of Psychiatry, Harvard Medical School, Boston, and The Cambridge Hospital. Address reprint requests to Dr. Westen, Department of Psychiatry, The Cambridge Hospital, 1493 Cambridge St., Cambridge, MA 02139; dw@wjh.harvard.edu (e-mail).

The author thanks Howard Shaffer for comments on earlier drafts of this article and Danny Koren for statistical advice and consultation.

constructs it purported to assess. An illustrative example is a study conducted by Skodol et al. (8), which showed only marginal associations among diagnoses made by the Structured Clinical Interview for DSM-III-R Personality Disorders (SCID-II) (9, 10), the Personality Disorders Examination (11), and the LEAD standard (longitudinal expert evaluation through use of all available data [12]). Equally problematic, but typical of studies correlating self-report with interview measures, was a study by Torgersen and Alnaeus (13) of the relation between diagnoses made by the self-report Millon Clinical Multiaxial Inventory (14) and an interview measure, the Structured Interview for DSM-III Personality Disorders (15). Correlations between dimensional measures of 11 personality disorders from the two instruments ranged from a high of 0.42 (avoidant) to a low of –0.05 (obsessive-compulsive), with most correlations in the range of 0.20–0.30. Self-report measures have tended to perform particularly poorly, although they may be useful as screening devices (16).

Perry's review (7) showed that the average kappa between any two instruments was only 0.25, with interviews fairing slightly better. Kappa (17) is a statistic for assessing whether two raters or measures categorize subjects or responses similarly, in this case, whether two instruments similarly categorize patients by diagnosis. Because kappa is roughly translatable into degree of variance explained, this means, as Perry pointed out, that a substantial part of the variance in personality disorder diagnoses made by axis II measures is likely to be error. The kappa is higher for some diagnoses, such as borderline and antisocial personality disorder, and considerable advances have been made in the study of these disorders by using these instruments (such as the recognition of the role of childhood sexual abuse in the etiology of borderline personality disorder [18]). In general, however, these instruments have not approached the standards of validity used in other areas of personality research.

The kappa for the simple presence/absence of a personality disorder across instruments is substantially higher than that for any specific disorder; that is, the various measures show far better agreement on whether a person has *some* personality disorder, and this has proven useful in predicting course of illness. The lack of discriminant validity (between disorders), however, remains a considerable problem; an axis I instrument would not be considered valid if it could diagnose *presence* of psychopathology but could not discriminate anorexia from substance abuse or schizophrenia. The picture is much more sanguine with respect to reliability than validity, although recent research assessing the latest version of the SCID-II (19) shows test-retest reliability well below acceptable levels for every personality disorder, even though the retest was within 2 weeks of the original administration of the interview. Further, most studies assessing test-retest reliability of existing measures with an interval of more than 2 weeks have found reliability to be unacceptably low, with kappas as low as 0.11 for some disorders (15, 19).

The validity problems of existing measures of personality disorders could reflect several causes. One possibility is that the measures have difficulty because DSM-IV itself does not coherently categorize disorders. Axis II instruments cannot establish criterion validity—predicting a criterion, such as other measures—if the DSM-IV personality disorders themselves lack construct validity. A second possibility is that we simply lack a gold standard against which to assess current measures (12), since clinical judgment itself is not reliable with respect to personality disorder diagnosis. A third possibility is that limits in the test-retest reliability of current measures produce a ceiling on interinstrument agreement, since the association between two unreliable instruments is likely to be low.

CURRENT RESEARCH PROCEDURES DIVERGE FROM CLINICAL INFERENCE PROCESSES

All of these issues probably contribute to the validity problems of current measures. A fourth possibility, however, is that the problems may stem in part from the fact that these instruments bear little resemblance to the way clinicians actually draw inferences about personality. Current measures—whether self-report questionnaires or semistructured interviews—share one essential design feature: they try to arrive at diagnoses primarily by asking patients direct questions derived from axis II criteria. In the 1980s this seemed a natural way to proceed, since studies relying on similar measures for assessing axis I disorders, such as the Schedule for Affective Disorders and Schizophrenia (20) and the Structured Clinical Interview for DSM-III-R (21, 22), were yielding useful data, for example, regarding prognosis and etiology, for many disorders.

What was never tested, however, was the extent to which clinicians use such questions when making axis II diagnoses. Asking a patient who has presented with depression, "Have you gained or lost much weight lately?" and "Do you find yourself crying a lot?," is common and intuitively sensible in clinical practice. Although patients' responses to questions such as these can be substantially affected by biases related to insight or motivation, these questions nonetheless provide a skeleton for assessment of disorders such as depression or bulimia, unless the patient has a reason to conceal symptoms.

Matters may be somewhat different for assessing disturbances in personality. For example, the Structured Interview for Personality Disorders (15) asks, "Have you ever been told that you seemed like a shallow or superficial kind of person?," to assess histrionic personality disorder. The SCID-II (9) asks, "Do you feel that your situation is so special that you require preferential treatment?," to assess narcissistic personality disorder. In each case the researchers constructing the instruments have taken special care to try to circumvent social desirability effects and defensiveness as much as possible—asking if "you have ever been told" instead of "are you," for example, or beginning with one or two open-ended questions such as, "How would you describe yourself as a person?" and

"What kinds of things have you done that other people might have found annoying?" (9). Nevertheless, one might question the utility of relying primarily on direct questions, at least for several of the personality disorders (in which lack of insight is diagnostic) or for many of the criteria for specific disorders (which may require judgments that patients lack the training, insight, or objectivity to make).

Indeed, several lines of evidence point to the limits of assessing personality processes primarily with direct questions (23). First, understanding people requires training. Psychiatrists would not need 3- to 5-year residencies if diagnoses or subtle psychological processes were so readily apparent to lay observers. Second, recent research documents a crucial distinction in the realms of memory, cognition, affect, and motivation between explicit (consciously accessible) and implicit (consciously inaccessible) processes. For example, memory researchers now distinguish between explicit and implicit memory (24). Explicit memory is conscious memory for facts and events. Implicit memory cannot be brought to mind consciously but is expressed in behavior. Priming studies and research with amnesic patients demonstrate that these systems can be entirely disconnected and are neurologically distinct, as when an amnesic patient with hippocampal damage has no conscious memory for a list of words but tends to complete word stems with words from the list when later asked simply to name the first word that comes to mind that begins with those letters (25). Just as people unconsciously use grammatical rules to construct sentences regardless of whether they can articulate the rules, they similarly use rules, for example, to respond to emotional cues, to guide their interpersonal behavior, and to make inferences about people's motives, to which they have no cognitive access.

Third, a considerable body of research has now documented that people do, indeed, use unconscious rules for transforming information defensively and that these processes considerably bias their answers to direct questions about themselves, particularly when the questions have implications for their self-esteem. Several independent literatures have produced evidence that discrepancies between what people consciously report feeling and what they show physiologically are predictable from reliable assessments of aspects of personality such as repressive coping style, attachment style, and defensiveness (26–28). For example, subjects who report an absence of psychological symptoms but whose early memories show signs of distress show a pattern of cardiac reactivity predictive of heart disease, and self-report lie scales (which are not included in personality disorder interviews) do not detect these individuals (28, 29).

THREE POTENTIAL PROBLEMS AND THREE CORRESPONDING HYPOTHESES

The primary aim of the present study was to examine the extent to which current personality disorder measures mirror clinical diagnostic processes. We identified three potential problems with existing measures and tested three corresponding hypotheses by surveying a study group of clinicians experienced in the treatment and supervision of treatment of personality disorders and then replicating the major findings with a random national sample of highly experienced psychiatrists and psychologists.

1. Instruments for assessing personality disorders may have transposed a method of assessing axis I disorders onto axis II diagnosis without careful enough attention to differences in methodology required by differences in the phenomena being assessed. Thus, we predicted that when clinicians assess personality pathology, they rely less on direct questions derived from diagnostic criteria than on a) patients' narrative descriptions of themselves, their past, and their interactions with others; and b) observation of the patient's behavior in the room, particularly with the clinician. We also hypothesized that clinicians consider direct questions more useful for diagnosing axis I than axis II disorders.

2. A patient who receives any personality disorder diagnosis through use of current instruments typically receives several. Skodol and colleagues (30, 31) have found that patients with personality disorders typically receive three to six axis II diagnoses if they receive any. Given that axis II includes only 10 categories, this suggests a problem with construct validity (that is, that the 10 categories may not represent coherent, differentiable constructs), discriminant validity (that the measures have trouble drawing distinctions that can be drawn), or both. Previous research (5) suggests that clinicians are more likely to prioritize personality disorder diagnoses, rather than giving multiple diagnoses. Thus, we hypothesized that a) when clinicians describe their own patients, the modal number of personality disorder diagnoses would be one, and b) clinicians would report prioritizing diagnoses rather than giving multiple diagnoses.

3. Axis II instruments, like axis II itself, may fail to assess maladaptive personality patterns that are not severe enough to meet axis II criteria but nevertheless bring patients in for treatment and attract therapeutic attention. Thus, we hypothesized that clinicians would report treating patients psychotherapeutically for enduring personality patterns that cannot be assigned to any axis II diagnosis, many of which are also not readily located on axis I.

METHOD

Subjects

Subjects for the first study were 52 clinical staff and faculty associated with Harvard Medical School at The Cambridge Hospital who responded to a survey. By virtue of the patient population at the hospital, which includes a heavy mix of patients with psychoses and personality disorders, all respondents have considerable experience with treatment or supervision of patients with axis II pathology. Most also have private practices that include patients with axis II disorders. The study group consisted of 17 psychiatrists, 32 psychologists, and three clinical social workers; mean length of posttraining experience was 9.80 years (SD=8.56, range=0–36).

TABLE 1. Importance of Five Methods for Diagnosing Personality Disorders and Degree Relied on by 51 Clinicians[a]

Method	Importance Rank[b] (N=30) Mean	SD	Reliance Rating[c] (N=51) Mean	SD
1. Asking direct questions derived from DSM-IV	3.37	0.93	5.31	1.71
2. Listening to the way patient describes interactions with significant others	1.20	0.48	1.10	0.41
3. Observing patient's behavior, including with you	1.87	0.57	1.22	0.50
4. Speaking with significant others	3.63	0.89	5.20	1.52
5. Administering questionnaires	4.67	0.48	6.51	0.86

[a]Data analyzed by repeated measures ANOVA, with Scheffé post hoc contrasts. The same pattern emerged with Wilks's lambda based on MANOVA with contrasts based on univariate F tests. Post hoc contrasts were all significant ($p<0.01$).
[b]Ranked 1–5; lower scores indicate higher ranking. Method 2<3<1, 4<5; $F=104.90$, $df=4, 116$, $p<0.0001$.
[c]Rated 1–7; lower scores indicate higher rating. Method 2, 3<1, 4<5; $F=272.92$, $df=4, 200$, $p<0.0001$.

Procedure

Subjects received a survey identified as designed to find out how clinicians think about personality disorders and how to diagnose them. The survey consisted of several items designed to test the three hypotheses.

1. Subjects were first presented with a table consisting of five methods that they were asked to rank in order of importance for diagnosing personality disorders. They were then asked to "rate each method (on a scale from 1 to 7) for the degree to which you rely on it in clinical practice to diagnose personality pathology, where 1 means 'I rely on it very much' and 7 means 'I rely on it very little'." The five methods were as follows: asking direct questions derived from DSM-IV axis II criteria, such as, "Do you think that it's not necessary to follow certain rules or conventions when they get in your way?" or "Do you feel that your situation is so special that you require preferential treatment?" (Sample items were taken from the SCID-II.); listening to the way the patient describes interactions with significant others from the past and present, such as how the patient perceives the self and others, whether the patient can tell coherent and sensible interpersonal narratives, and how the patient describes emotionally charged material; observing the patient's behavior in the room, including his or her way of interacting with the clinician; speaking with the patient's significant others; and giving the patient self-report questionnaires.

Next, subjects were asked to rate each of the axis II diagnoses and four axis I diagnoses (schizophrenia, major depression, panic disorder, and anorexia nervosa) for "the extent to which direct questions derived from DSM-IV criteria (e.g., 'Do you feel that your situation is special so that you require preferential treatment?' or 'Do you hear voices') are useful in diagnosis" (rating of 1=very useful, 7=not very useful). The two examples (one for axis I and one for axis II) were chosen to be relatively comparable, in that clinicians might be likely to consider the possibility of denial in both, since some narcissists might deny entitlement, just as some schizophrenic patients might deny psychotic symptoms.

2. To see how many concurrent personality disorder diagnoses clinicians actually give in practice, and whether they prioritize when making axis II diagnoses, clinicians were asked to respond to two tasks. They were asked to list the initials of up to five patients they currently treated who had personality disorders, to list all axis II disorders for which the patients fully met criteria, and to rate on a scale from 1 to 7 the degree to which each diagnosis adequately described the patient's personality pathology. Subjects were asked for initials so they would be likely to use actual patients rather than drawing on prototypes. Next, subjects were asked the following question: "If you give a patient more than one personality disorder diagnosis, do you usually consider one diagnosis to be primary? Yes, No, NA—I rarely give multiple personality disorder diagnoses."

3. Finally, subjects were asked the following two questions about personality problems that do not meet axis II criteria but that nevertheless may bring patients to treatment or receive clinical attention in psychotherapy: "Do patients ever come to you for treatment of 'neurotic' personality patterns that are not severe enough to meet axis II criteria? Yes, No, NA—I am primarily a psychopharmacologist." If subjects answered in the affirmative, they were asked to give specific examples of the kinds of problems with which patients presented. Finally, respondents were asked, "Among your psychotherapy patients who *do not* have axis II diagnoses, what percent have 'neurotic' personality patterns or styles that you address in treatment?"

RESULTS

In all cases a rating of 1 refers to clinicians' judgments of high importance, high reliance in clinical practice, or high confidence (depending on the question), whereas a rating of 7 refers to low importance, low reliance in clinical practice, or no confidence.

Hypothesis 1: Perceived Clinical Utility of Various Diagnostic Methods

The first two questions concerned the importance that clinicians ascribe to different methods of diagnosing personality pathology (ranking five methods) and the degree to which they report relying on each method clinically. These two questions provided slightly different ways of measuring the same issue, so that convergence between them would lead to greater confidence in the results. Data reported in table 1 reflect repeated measures analysis of variance (ANOVA), although multivariate ANOVA (MANOVA) produced the same results for both the overall differences among the methods and contrasts among particular methods. As can be seen from table 1, the pattern was quite clear from both rankings and ratings: clinicians valued, and relied primarily on, the way patients describe interactions with significant others and their behavior in the room, particularly with the interviewer, and they found direct questions derived from axis II, information from informants, and questionnaires much less useful in making axis II diagnoses. Post hoc pairwise contrasts comparing the five methods produced significant findings through use of both ratings and rankings, and the major hypotheses, that clinicians would find direct questions less useful than listening to narratives and observing the patient's behavior in the room, were all supported through use of both ratings and rankings ($p<0.0001$) despite the small group sizes.

To test the hypothesis that clinicians find direct questions derived from DSM criteria less useful for axis II than axis I diagnoses, we averaged the ratings of the usefulness of asking DSM-IV-derived questions for each of the personality disorders into an axis II rating and compared this with the average of the similar ratings for the four axis I diagnoses. As predicted, the ratings were vastly different: axis I, mean=1.75 (SD=1.17); axis II, mean=4.10 (SD=

1.30) (t=12.18, df=49, p<0.0001). Repeated measures ANOVAs comparing different disorders within each axis were not significant.

Hypothesis 2: Clinicians Tend to Give One Diagnosis and to Prioritize Diagnoses if Patients Meet Criteria for Multiple Diagnoses

Collectively, the 52 clinicians described 162 patients, for a mean of 3.12 patients per clinician. For each of these patients, the mean number of personality disorder diagnoses was 1.28 (SD=59, range=1–3), with a mode of 1. When asked whether they typically consider one diagnosis primary, 44.0% said yes, 14.0% said no, and 42.0% said that they too rarely give more than one axis II diagnosis to answer the question. Thus, 86.0% of the clinicians give patients with axis II disorders one primary diagnosis.

Hypothesis 3: Clinicians Treat Patients for Maladaptive Personality Patterns Not Diagnosable on Axis II

A total of 86.5% of clinicians reported that patients sometimes present for treatment of personality problems that cannot be diagnosed on axis II. (The remainder of subjects either answered in the negative or said that the question was inapplicable because they were primarily psychopharmacologists.) The clinicians reported that they treated a mean of 60.8% (SD=32.2%) of their psychotherapy patients *without* axis II disorders for personality problems.

When asked for specific examples of the types of personality problems that bring their patients without axis II disorders in for treatment, clinicians reported a wide range of difficulties, many of which fall into neither axis I nor axis II. Four categories were particularly common. First, of the 40 clinicians who provided descriptions of the neurotic personality patterns that they typically treat, fully half (20 subjects) mentioned problems of intimacy, relatedness, and commitment. It is important to note that none of these problems can be reduced to an axis I syndrome. Several pointed to related interpersonal issues such as lack of assertiveness, "avoidant tendencies," self-defeating behavior, authority problems, shyness, passive-aggressive traits, conflicted identification with a parent, unresolved grief, and problems with separation or rejection. Only some of these resemble axis I or II criteria, and those that do cannot be *assumed* to reflect a subsyndromal disorder. For example, problems with separation or rejection occur in some relatively high-functioning patients who do not have prominent anxiety symptoms or borderline features, which are the only places such problems can be diagnosed. Second, 45.0% described problems with work, including work inhibitions, chronic dissatisfaction, underachievement, and lack of direction in life. None of these problems can be located on either axis. A third category was depressive-proneness or characterological depression that most clinicians reported was not severe enough to meet either axis I or axis II criteria (35.0%). One could argue that these are subclinical manifestations of a mood disorder, although they are just as easily conceptualized as personality disturbances and hence should be diagnosed on a personality axis. Fourth, 30.0% said that their patients came for treatment of low self-esteem, low self-confidence, feelings of inadequacy, or feelings of unlovability that meet neither axis I nor axis II criteria. Several other categories of symptomatic personality patterns are worth noting. Of the respondents, 22.5% described obsessional patterns such as affective constriction, trouble making decisions, overcontrol, rigidity, and intellectualization as problems, and they frequently noted that patients with these patterns were often high functioning (unlike the prototypical obsessive-compulsive patient, as described by axis II). Twenty percent described narcissistic problems not severe enough to warrant an axis II diagnosis, such as fluctuating between devaluing and idealizing the self and others. Fifteen percent described chronic problems with anxiety or anxious apprehension. Other problems described by more than one clinician included problems with affects other than depression and anxiety (notably anger and guilt), impulsivity, and perfectionism.

A Replication

Because the findings of this study (and particularly the first hypothesis, regarding the methods clinicians use and trust in making personality disorder diagnoses) could be biased by the use of a study group of clinicians from a single institution, we conducted a replication study using a random national sample of licensed clinicians. As part of a study aimed at validating a new instrument for assessing personality pathology, we contacted 3,000 psychiatrists from the register of the American Psychiatric Association who indicated an interest in personality disorders and 4,000 psychologists from the American Psychological Association who were selected from the three divisions that draw clinicians (the divisions of clinical psychology, psychotherapy, and psychoanalysis). In both cases, the selection procedure included a computer search to exclude clinicians with less than 3 years' posttraining practice. Clinicians were asked, among other things, to check off from a list of axis II disorders all categories from which they had treated a patient in the last 6 months and to check off their primary theoretical orientation (biological, psychodynamic, cognitive-behavioral, systemic, or eclectic). They were also asked the same questions used to test the first hypothesis from the study described earlier, namely, those items that asked them to rate the usefulness of different sources of data for making axis I and axis II diagnoses, such as asking direct questions compared with listening to the patient's narrative descriptions of events.

At the time these data were analyzed, a total of 1,901 respondents (1,305 psychologists and 596 psychiatrists) had returned the survey. (Approximately 50 respondents provided incomplete data, so that the numbers in the analyses reported later in this article are

TABLE 2. Ratings of Usefulness of Various Methods for Diagnosing Personality Disorders by 1,827 Psychiatrists and Psychologists, by Theoretical Orientation

	Method												
	Direct Questions		Narratives		Observe Behavior		Informant		Self-Report		Analysis[a]		
Clinician Orientation	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	F	df	p
Biological (N=54)	4.20	1.94	1.69	1.19	1.69	1.04	3.22	1.99	5.46	2.07	55.87	4, 212	<0.0001
Psychodynamic (N=824)	5.11	1.58	1.20	0.63	1.31	0.76	5.28	1.65	6.09	1.48	3,079.09	4, 3,292	<0.0001
Cognitive-behavioral (N=292)	4.29	1.78	1.75	1.12	1.93	1.24	3.81	1.66	3.89	1.99	180.90	4, 1,164	<0.0001
Family systems (N=25)	5.12	1.56	1.50	0.58	1.42	0.64	3.23	1.21	4.24	1.85	39.23	4, 96	<0.0001
Eclectic (N=632)	4.62	1.92	1.47	0.92	1.60	0.98	3.98	1.77	5.22	1.93	846.77	4, 2,524	<0.0001
Total (N=1,827)	4.96	1.82	1.40	0.87	1.52	0.96	4.51	1.84	5.39	1.92	3,132.89	4, 7,336	<0.0001

[a]Data analyzed by repeated measures ANOVA. MANOVA with univariate F tests produced the same pattern of findings. All contrasts between the methods narratives/observe behavior and the other three methods were significant (p<0.0001).

slightly smaller.) The respondents were a highly experienced group: the mean length of postresidency or postdoctoral training was 18.18 years (SD=9.56). Several had multiple institutional affiliations: 613 worked in hospitals at least part time, 384 worked in clinics, 1,573 had private practices, and 214 worked in forensic settings. When asked their primary theoretical orientation, 56 identified themselves as biological, 838 as psychodynamic, 300 as cognitive-behavioral, 36 as systemic, and 639 as eclectic.

As can be seen from table 2, the findings from the first study replicated for the entire sample and for clinicians of every theoretical orientation. Once again, we used repeated measures ANOVA, but MANOVA with pairwise contrasts produced the same pattern of findings. Whether biological, cognitive-behavioral, psychodynamic, systemic, or eclectic in orientation, clinicians did not find direct questions as useful for assessing axis II pathology as listening to patients' narratives and observing their behavior in the room. In contrast, they found such questions of substantially greater utility for assessing axis I disorders. The comparison of ratings of the usefulness of asking direct questions regarding axis II disorders and axis I disorders (again calculated as the mean across four diagnoses) was highly significant (mean=4.96, SD=1.82, versus mean=2.67, SD=1.70) (t=48.23, df=1,835, p<0.0001).

Also of interest from this national survey were the percentages of clinicians who reported treating various disorders, some of which have now been removed from the text and the appendices to axis II on the basis of research findings (e.g., passive-aggressive personality disorder) or political considerations (e.g., self-defeating personality disorder). Among disorders currently on axis II, schizotypal was the least widely treated by clinicians in this sample (32.0% reported currently treating at least one schizotypal patient), and borderline was the highest (85.0%). The others hovered around 50%, except for narcissistic (76.0%), dependent (72.0%), and obsessive-compulsive (68.0%). It is of interest that depressive personality disorder, which is not in DSM-IV, was eclipsed only by borderline personality disorder (at 77.0%). Passive-aggressive (58.0%) and self-defeating (52.0%) personality disorders were also common, unlike sadistic personality disorder (13.0%).

DISCUSSION

The data point to several conclusions. First, when researchers began creating structured interviews for personality disorders on the model of axis I interviews, they diverged from clinical practice in a way that was not the case with research instruments developed to assess axis I. Rather than relying primarily on direct questions to assess maladaptive personality patterns, clinicians attend carefully to patients' narratives, particularly those involving interpersonal interactions, and to the way patients interact with the interviewer. It is striking that these findings were robust across theoretical orientations: no matter what their theoretical orientation, clinicians found direct questions of limited utility for assessing personality disorders. (Although no differences emerged across personality disorder diagnoses in reported utility of direct questions, it may be that such differences would emerge across *criteria,* such as asking about friendship patterns compared with quality of emotional experience in schizotypal patients, who may know that they have few friends but not that their bland affect or emotional expression is peculiar. The utility of direct questions may vary somewhat across axis I criteria as well and would be worthy of empirical attention, although the reported utility of direct questions clearly differs between axis I and axis II.)

A potential rejoinder by advocates of these instruments would be that clinicians, too, lack the ability to make reliable or valid judgments, and hence the divergence from clinical diagnostic and inference processes is not particularly troubling. The problem with this argument is that personality disorders were discovered and initially classified on the basis of clinical observation, through use of methods that apparently diverge substantially from those now used in research guiding successive revisions of axis II. The most obvious way to assess personality characteristics would certainly be to ask people if they have them, but if clinicians have not

gravitated toward this method over the last 50 years, we should carefully consider why they have not.

Indeed, examination of the evolution of axis II instruments suggests that the divergence from clinical methods may have been an unintended consequence of what began as a sound decision from a research standpoint. The first personality disorder instrument, the Diagnostic Interview for Borderline Patients (32), required 90 minutes to administer to assess a single disorder and did so because it required substantial clinical probing and acumen. Over time, researchers recognized the importance of assessing multiple personality disorder diagnoses simultaneously, since comorbidity in personality disorder diagnoses meant that a study of borderline personality disorder might actually include many subjects who could equally be categorized as histrionic or schizotypal. This led to the development of instruments that attempt to assess all the personality disorders in a single interview. To accomplish this within a reasonable period of time (typically 1–3 hours, or 30 minutes following administration of a questionnaire for the SCID-II), researchers came to rely increasingly on direct questions.

This heavy reliance of research instruments on a method that clinicians find of limited utility in assessing personality disorders has begun to affect the disorders and criteria included in DSM-IV in ways that need to be carefully considered. For example, passive-aggressive personality disorder was eliminated from DSM-IV, in part because of its apparent rarity on the basis of current interviews (J. Gunderson, personal communication, August 1995). In our national sample of clinicians, however, 58.0% reported currently treating at least one patient who fully met criteria for this disorder. This is a higher percentage than that for five of the 10 disorders now represented on axis II. One possibility, of course, is that clinicians may see passive aggression where it does not exist. An alternative explanation is that passive-aggressive personality disorder—like several other personality disorders, such as narcissistic, histrionic, and schizoid—cannot be adequately assessed through direct questioning because a defining characteristic of the disorder is substantial self-deception or lack of insight.

Research diagnoses diverge from clinical diagnoses for axis II disorders in a second way: they diagnose multiple disorders where clinicians do not. Three explanations could account for this divergence. First, clinicians may fail to recognize comorbidity, in contrast to structured instruments, which may lead to more "evenly hovering attention" across potential diagnoses. A second alternative faults the instruments rather than clinicians: either the questions they ask or the criteria they assess (or both) may not allow for discrimination among disorders. In clinical practice, the modal patient with a personality disorder receives only one axis II diagnosis, and the mean patient receives only 1.27. Although comorbidity is certainly common on both axis I and axis II, and some axis I disorders may not be as discrete as traditionally supposed (e.g., mood and anxiety disorders, or bipolar disorder and paranoid schizophrenia), in more cases than not, competent clinicians and researchers can accurately distinguish a patient with schizophrenia from one with dysthymia or major depression. The same cannot be said for the distinction between schizotypal and borderline personality disorders, which appear clinically quite dissimilar but show substantial comorbidity on research instruments. A third possibility is that the instruments tend to diagnose multiple disorders because they cannot prioritize diagnoses. The data from this study point to a substantial contrast between clinical and research diagnosis but cannot adjudicate among these alternative explanations.

Some of these considerations probably apply to axis I as well, although this is beyond the scope of the present article and should be the target of further research. For example, assessing most axis I syndromes requires inference and clinical skill (especially when a patient with a disorder such as anorexia, schizophrenia, or bipolar disorder denies symptoms), although the clinicians we surveyed clearly found direct questions less problematic for axis I than axis II disorders. Similarly, record reviews at major psychiatric hospitals show that most clinicians do not give patients multiple axis I diagnoses, either (33), in contrast to research instruments. This divergence may be somewhat less problematic for axis I than axis II, however, since axis I includes dozens of disorders rather than 10, so that the difference between diagnosing one versus two or three disorders may have less bearing on discriminant validity.

Current instruments diverge from clinical practice in a third way that mirrors a flaw in axis II itself: they fail to include personality patterns that bring people to treatment and require clinical intervention but are not severe enough to warrant an axis II diagnosis. Many of these symptoms and patterns, such as recurring problems with intimacy, work, and self-esteem, are also not part of any axis I syndrome. These problems are clearly aspects of personality—that is, they are enduring patterns of thought, feeling, motivation, and behavior that occur over time—such as authority problems, difficulty with self-assertion, and sensitivity to rejection or abandonment in patients who clearly do not have borderline pathology (which is the only axis II disorder that includes fear of abandonment as a symptom). By including only a small set of categories and limiting these to severe forms of personality disturbance, DSM-IV, and the instruments derived from it, has effectively left out a large spectrum of psychopathological conditions, once called "neurotic," that need to be assessed by any instrument purporting to measure personality pathology and should be reconsidered for inclusion in DSM-V.

CONCLUSIONS

Because clinician diagnoses are themselves often unreliable, reverting to unstructured clinical observation is not a solution. A potential solution, however, may lie in the distinction between two processes that clinicians use in making an axis II diagnosis, one of which appears

to be more reliable than the other. The first process entails, as the respondents to this study told us, observing patients' interactions in the consulting room, listening to their narratives about their lives, and drawing inferences about their characteristic behavior, conscious coping strategies, unconscious affect-regulatory procedures (defenses), cognitive patterns, wishes, fears, values, and affective propensities. A growing body of research suggests that through use of psychometrically sound instruments, clinicians can, in fact, make such inferences reliably (34–39; J. Shedler, D. Westen, unpublished data, 1996), particularly if statements that they are rating or ranking are written in plain language with minimal jargon.

The second process in making an axis II diagnosis is to apply an algorithm to combine those dozens of observations into a diagnosis. DSM-IV forces a particular kind of algorithm: whether the patient meets five of eight criteria, and so forth. In actual practice, however, clinicians probably use the same mix of algorithms used by people in all categorization tasks (40–42). That is, much of the time they match their pattern of observations against a prototype or exemplar of a category, and if the match is good, they conclude that the particular instance is a member of the category. At other times, particularly if some piece of data seems anomalous, they consult a list of defining features, as in DSM-IV.

Application of these algorithms is probably much less reliable than the observations that underlie them, particularly when the diagnostic categories themselves do not lie on solid empirical ground. A potential solution is to use procedures developed by personality researchers for measuring complex personality processes, which *statistically* compare the observed pattern of personality attributes of a given patient with the patterns found among particular groups (37, 43; J. Shedler, D. Westen, unpublished data, 1996). Thus, if a cluster of patients empirically exists who share a common set of characteristics resembling the antisocial diagnosis, the profiles of a large number of such patients are aggregated to form a prototype of that diagnosis by using an instrument that assesses a wide array of personality characteristics. To test whether a given patient meets criteria for that diagnosis for research purposes, a clinician or researcher interviews the patient, attending carefully, among other things, to the patient's narrative descriptions of salient interpersonal encounters, and then describes the patient by using the personality descriptors that comprise the items in the instrument. Then, rather than intuitively combining these observations into a diagnosis, as in clinical diagnosis, or counting up criteria based on direct questions, as in research diagnosis, the clinician or researcher statistically *correlates* the patient's profile across these items with the empirically observed prototype (that is, the average profile of an antisocial patient on the instrument) to assess the degree of match between the patient and the prototype.

A virtue of this method is that it can yield dimensional diagnoses determined by simple correlation coefficient (e.g., the correlation between the patient's profile and the antisocial profile is 0.63), categorical diagnoses (based on a cutoff score), or a combination of the two (e.g., "antisocial personality disorder with borderline features") based on a combination of cutoffs and dimensional scores. Another virtue is that the items in the instrument and the prototype profiles need not be limited to severe personality disorders and can, in fact, assess the spectrum of personality processes, ranging from relatively healthy to relatively disturbed. Preliminary research using procedures of this sort has yielded promising results (37, 38; J. Shedler, D. Westen, unpublished data, 1996) and suggests that clinical observation and research observation need not be so divergent with respect to the diagnosis of personality disorders. Indeed, the assessment and categorization of personality disorders are likely to be enhanced substantially if instruments demonstrate both *clinical* and *empirical validity*.

REFERENCES

1. Clark L: Resolving taxonomic issues in personality disorders: the value of larger scale analyses of symptom data. J Personality Disorders 1992; 6:360–376
2. Jackson D, Livesley WJ: Possible contributions from personality assessment to the classification of personality disorder, in The DSM-IV Personality Disorders. Edited by Livesley WJ. New York, Guilford Press, 1995, pp 459–481
3. Morey LC: Personality disorders in DSM-III and DSM-III-R: convergence, coverage, and internal consistency. Am J Psychiatry 1988; 145:573–577
4. Oldham JM, Skodol AE, Kellman HD, Hyler SE, Rosnick L, Davies M: Diagnosis of DSM-III-R personality disorders by two structured interviews: patterns of comorbidity. Am J Psychiatry 1992; 149:213–220
5. Widiger T, Frances A: Towards a dimensional model for the personality disorders, in Personality Disorders and the Five-Factor Model of Personality. Edited by Costa P, Widiger T. Washington, DC, American Psychological Association, 1994, pp 19–39
6. Gunderson J: Diagnostic controversies, in American Psychiatric Press Review of Psychiatry, vol 11. Edited by Tasman A, Riba M. Washington, DC, American Psychiatric Press, 1992, pp 9–24
7. Perry JC: Problems and considerations in the valid assessment of personality disorders. Am J Psychiatry 1992; 149:1645–1653
8. Skodol A, Oldham J, Rosnick L, Kellman D, Hyler S: Diagnosis of DSM-III-R personality disorders: a comparison of two structured interviews. Int J Methods in Psychiatr Res 1991; 1:13–26
9. First M, Spitzer R, Gibbon M, Williams J: The Structured Clinical Interview for DSM-III-R Personality Disorders (SCID-II), part I: description. J Personality Disorders 1995; 9:83–91
10. Spitzer RL, Williams JBW, Gibbon M: Structured Clinical Interview for DSM-III-R Personality Disorders (SCID-II). New York, New York State Psychiatric Institute, Biometrics Research, 1987
11. Loranger AW, Susman VL, Oldham JM, Russakoff M: The Personality Disorder Examination (PDE) Manual. Yonkers, NY, DV Communications, 1988
12. Spitzer RL: Psychiatric diagnosis: are clinicians still necessary? Compr Psychiatry 1983; 24:399–411
13. Torgersen S, Alnaeus R: The relationship between the MCMI personality scales and DSM-III, Axis II. J Pers Assess 1990; 55: 698–707
14. Millon T: Millon Clinical Multiaxial Inventory, 3rd ed. Minneapolis, National Computer Systems, 1984
15. Stangl D, Pfohl B, Zimmerman M: A structured interview for DSM-III personality disorders: a preliminary report. Arch Gen Psychiatry 1985; 42:591–596
16. Hyler S, Skodol AE, Oldham JM, Kellman HD, Doidge N: Validity of the Personality Diagnostic Questionnaire, Revised: a replication in an outpatient sample. Compr Psychiatry 1992; 33: 73–77

17. Cohen J: A coefficient of agreement for nominal scales. Educational and Psychol Measurement 1960; 20:37–46
18. Zanarini M (ed): The Role of Sexual Abuse in the Etiology of Borderline Personality Disorder. Washington, DC, American Psychiatric Press, 1995
19. First M, Spitzer R, Gibbon M, Williams J, Davies JB, Howes M, Kane J, Pope H, Rounsaville B: Structured Clinical Interview for DSM-III-R Personality Disorders (SCID-II), part II: multi-site test-retest reliability study. J Personality Disorders 1995; 9:92–104
20. Endicott J, Spitzer RL: A diagnostic interview: the Schedule for Affective Disorders and Schizophrenia. Arch Gen Psychiatry 1978; 35:837–844
21. Spitzer RL, Williams JBW, Gibbon M, First MB: User's Guide for the Structured Clinical Interview for DSM-III-R (SCID). Washington, DC, American Psychiatric Press, 1990
22. Williams JBW, Gibbon M, First MB, Spitzer RL, Davies M, Borus J, Howes MJ, Kane J, Pope HG, Rounsaville B, Wittchen H-U: The Structured Clinical Interview for DSM-III-R (SCID), II: multisite test-retest reliability. Arch Gen Psychiatry 1992; 49:630–636
23. Westen D: A clinical-empirical model of personality: life after the Mischelian ice age and the NEO-lithic era. J Pers 1995; 63:495–524
24. Schacter DL: Understanding implicit memory: a cognitive neuroscience approach. Am Psychol 1992; 47:559–569
25. Squire LR: Memory and Brain. New York, Oxford University Press, 1987
26. Weinberger DA: The construct validity of the repressive coping style, in Repression and Dissociation: Implications for Personality Theory, Psychopathology, and Health. Edited by Singer JL. Chicago, University of Chicago Press, 1990, pp 337–386
27. Dozier M, Kobak R: Psychophysiology in attachment interviews: converging evidence for deactivating strategies. Child Dev 1992; 63:1473–1480
28. Shedler J, Mayman M, Manis M: The illusion of mental health. Am Psychol 1993; 48:1117–1131
29. Shedler J, Mayman M, Manis M: More illusions. Am Psychol 1994; 49:974–976
30. Skodol A, Rosnick L, Kellman H, Oldham J, Hyler S: Development of a procedure for validating structured assessments of Axis II, in Personality Disorders: New Perspectives on Diagnostic Validity. Edited by Oldham J. Washington, DC, American Psychiatric Press, 1990, pp 47–70
31. Skodol AE, Rosnick L, Kellman D, Oldham JM, Hyler SE: Validating structured DSM-III-R personality disorder assessments with longitudinal data. Am J Psychiatry 1988; 145:1297–1299
32. Gunderson JG, Kolb JE, Austin V: The Diagnostic Interview for Borderline Patients. Am J Psychiatry 1981; 138:896–903
33. Mezzich J, Goodpastor W, Mezzich AC: Structural issues in diagnosis, in Issues in Diagnostic Research. Edited by Last CG, Hersen M. New York, Plenum, 1987, pp 87–98
34. Vaillant G (ed): Ego Mechanisms of Defense: A Guide for Clinicians and Researchers. Washington, DC, American Psychiatric Press, 1992
35. Leigh J, Westen D, Barends A, Mendel M: Assessing complexity of representations of people from TAT and interview data. J Pers 1992; 60:809–837
36. Westen D: Social cognition and object relations. Psychol Bull 1991; 109:429–455
37. Westen D, Muderrisoglu S, Fowler C, Shedler J, Koren D: Affect regulation and affective experience: individual differences, group differences, and measurement using a Q-sort procedure. J Consult Clin Psychol (in press)
38. Westen D: A model and a method for uncovering the nomothetic from the idiographic: an alternative to the five-factor model? J Res Personality 1996; 60:499–413
39. Jacobson W, Cooper AM: Psychodynamic diagnosis in the era of the current DSMs, in Psychodynamic Treatment Research: A Handbook for Clinical Practice. Edited by Miller NE, Luborsky L, Barber JP, Docherty JP. New York, Basic Books, 1993, pp 109–126
40. Holyoak K, Spellman B: Thinking. Ann Rev Psychol 1993; 44: 265–315
41. Rosch E: Principles of categorization, in Cognition and Categorization. Edited by Rosch E, Lloyd BB. New York, John Wiley & Sons, 1978
42. Cantor N, Genero N: Psychiatric diagnosis and natural categorization: a close analogy, in Contemporary Directions in Psychopathology: Toward the DSM-IV. Edited by Millon T, Klerman G. New York, Guilford Press, 1986, pp 233–256
43. Block J: The Q-Sort Method in Personality Assessment and Psychiatric Research. Palo Alto, Calif, Consulting Psychologists Press, 1978