Journal of Abnormal Psychology
2007, Vol. 116, No. 4, 810–822

0021-843X/07/$12.00 DOI: 10.1037/0021-843X.116.4.810

Personality Diagnosis With the Shedler-Westen Assessment Procedure (SWAP): Integrating Clinical and Statistical Measurement and Prediction

Drew Westen
Emory University

Jonathan Shedler
University of Colorado Health Sciences Center

This article describes the Shedler-Westen Assessment Procedure (SWAP), a personality assessment instrument intended for use by clinically experienced interviewers, designed to maximize both psychometric precision and clinical relevance. The article focuses on the latest edition of the instrument, the SWAP–II; its use in 2 recently completed large-sample projects; and the ways in which data from these projects are being used to revise and refine concepts of personality pathology and taxonomy. The article first details the development of the SWAP and its psychometric rationale. It then examines the use of SWAP data for purposes of (a) improving diagnostic criteria within the framework of the existing *Diagnostic and Statistical Manual of Mental Disorders* taxonomy, (b) developing a new classification of personality pathology based on empirically identified diagnostic groupings, and (c) identifying trait dimensions relevant to understanding personality syndromes and disorders. Finally, the article discusses future research directions and challenges.

Keywords: personality, assessment, diagnosis, taxonomy, clinical judgment

This article describes the Shedler-Westen Assessment Procedure (SWAP), a personality assessment instrument developed for use by clinically experienced interviewers, designed to maximize both psychometric precision and clinical utility. This article focuses primarily on taxonomic and psychometric issues; for a more general introduction to the SWAP and an overview of prior SWAP research, see Shedler and Westen (2007). The SWAP has been used to (a) refine and dimensionalize existing *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.; *DSM–IV*; American Psychiatric Association, 1994) diagnostic categories and criteria; (b) empirically identify diagnostic groupings without presupposing the *DSM–IV* typology of personality disorders (PDs); and (c) identify factors or trait dimensions relevant to describing personality pathology. We demonstrated these applications with earlier versions of the SWAP instruments (e.g., Shedler & Westen, 2004a, 2004b; Westen, Dutra, & Shedler, 2005; Westen & Shedler, 1999a, 1999b; Westen, Shedler, Durrett, Glass, & Martens, 2003). Here we focus on the latest editions of the instrument, the SWAP–II and SWAP–II–A (for adolescents), and describe findings from two large, recently completed projects.

We begin by describing the SWAP and its rationale and psychometric properties. We then review evidence concerning reliability and validity. We next describe the "virtual field trial" methodology employed in our recent projects and discuss findings concerning diagnosis, taxonomy, and derivation of trait dimensions. Finally, we address concerns and remaining challenges for the SWAP research program.

The SWAP–II and SWAP–II–A

The SWAP–II is based on the Q-sort method, which has been employed for many years in the study of both normal personality (e.g., Block, 1961/1978; Shedler & Block, 1990) and pathological personality (e.g., Westen & Shedler, 1999a, 1999b). The SWAP–II is a set of personality descriptive statements (items), each of which may describe a given patient well, somewhat, or not at all. A clinical assessor sorts the statements into eight categories based on the degree to which the statements describe the patient, from 7 (*highly descriptive*) to 0 (*not descriptive*). We developed a systematic Clinical Diagnostic Interview (CDI; Westen, 2002; Westen & Muderrisoglu, 2003, 2006)[1] that can be administered in approximately 2½ hr and yields sufficient patient information to score the SWAP reliably and validly. The interview can be used in either clinical or research contexts. When the interview is not used, clinicians can score the SWAP after 6 or more clinical contact hours with a patient (the lower limit we specify in our research protocols).

The distribution of scores is *fixed*, meaning that the assessor must assign a specified number of items to each score category (e.g., exactly eight items receive scores of 7). The original format of Q-sort instruments was on index cards, but most assessors now prefer software-based methods (a Web-based program that allows electronic sorting of virtual cards can be previewed at

This research was supported by National Institute of Mental Health Grants MH62377 and MH62378.

Correspondence concerning this article should be addressed to Drew Westen, Department of Psychology, and Department of Psychiatry and Behavioral Sciences, Emory University, 532 N. Kilgo Circle, Atlanta, GA 30322, or to Jonathan Shedler, Department of Psychiatry, University of Colorado Health Sciences Center, University North Pavilion, 4455 East 12th Avenue, A011-99, Denver, CO 80220. E-mail: dwesten@emory.edu or jonathan@shedler.com

[1] Available to mental health professionals to download at http://www.psychsystems.net

www.SWAPassessment.org). A detailed discussion of Q-sort methodology is beyond the scope of this article. Block (1961/1978) described its psychometric rationale in detail, and we refer the interested reader to his classic text. However, one important advantage of the Q-sort method is that it minimizes error variance or "noise" due to rater effects. With standard rating scales, some raters naturally gravitate toward extreme values, others toward moderate values, and some use the entire scale range. Thus, differences in scores reflect not only personality differences between the individuals assessed but also differences in the "calibration" of the raters. Raters may also use different implicit norms when rating items (e.g., how hostile is the patient, relative to *whom*?). It is therefore possible that two assessors could agree perfectly regarding their actual observations yet produce very different scale ratings.

The Q-sort, with its fixed score distribution, minimizes these sources of error variance by "calibrating" assessors and ensuring that different assessors assign scores with the same frequency. Thus, when any assessor assigns a score of 7 to a SWAP item, the meaning is always the same: Relative to the 200 items in this item set, the item is among the top 8 that are most defining of the patient's personality. Although the scoring procedure is ipsative, the resulting data can be treated like normative data (e.g., for measuring individual differences) and have proven valid for this purpose, as demonstrated, for example, by consistently high criterion validity with appropriate criterion variables (see below). One could, with some cleverness, create a contrived example to illustrate a hypothetical situation wherein Q-sort scores might in principle become problematic when interpreted normatively. However, there is no evidence that such situations occur in real-world use. The question of the validity of Q-sort scores used in this manner is an empirical one, appropriately addressed (as with any other psychometric measure) by data, not armchair analysis.

By minimizing error variance as described previously, the Q-sort method maximizes the opportunity to observe statistical relations where they exist but does not artifactually inflate reliability or validity coefficients.[2] The method also minimizes possible clinician biases (e.g., confirmation biases, primacy, recency) by ensuring that assessors attend systematically to all constructs subsumed by the item set. Essentially, an assessor must not only systematically entertain 200 hypotheses about personality processes but also compare and contrast those hypotheses (because scoring is ipsative). Obviously, this method bears little relation to the kind of unsystematic, free-form case descriptions clinicians might otherwise provide, and that have been criticized, appropriately, for lack of reliability and validity. Historically, Q-sort methods have been especially useful in capturing and quantifying information about subtle aspects of personality not readily assessed via self-report (e.g., Colvin, Block, & Funder, 1995; Shedler & Block, 1990; Westen & Muderrisoglu, 2003, 2006).

We used multiple methods to ensure that the fixed score distribution would be appropriate for describing most patients and not an arbitrary imposition. We observed how clinicians rated SWAP items when we did not impose a fixed distribution, to determine the distribution clinicians used naturally. We also included a sufficiently broad range of item content (including more than 20 items that assess psychological strengths, plus a wide range of personality characteristics not associated with PDs; see Blagov, Bradley, & Westen, 2007) to ensure that there would always be enough items that "belong" (normatively) in the highest or most descriptive score categories.

This last point is crucial, and failure to understand it could lead to erroneous conclusions about the effects of a fixed distribution. If a fixed distribution were imposed on an instrument with more restricted item content, designed to assess a single construct (e.g., the Beck Depression Inventory II; Beck, Steer, & Brown, 1996]), the impact of the fixed distribution would be to render the scores normatively meaningless. The fixed distribution would presuppose that all patients had the same level of pathology, and the instrument would therefore underestimate depression in severely depressed patients and overestimate it in nondepressed patients. This does not apply to an omnibus instrument such as the SWAP, with item content covering a very broad range of psychological constructs spanning the spectrum of functioning from exceptionally psychologically healthy to severely disturbed. As long as there are enough items in the item set that normatively belong in the higher score categories—and we have yet to encounter a situation with the SWAP in which this was not so—then the fixed distribution will permit a normatively accurate portrayal of the individual.

Prior Q-sort instruments have treated items as bipolar dimensions (*extremely characteristic* to *extremely uncharacteristic*) and have used quasi-normal score distributions in which middle scores indicated neutrality on the dimension (e.g., Block, 1971; Shedler & Block, 1990). An innovation of the SWAP is that all items are written to assess unipolar constructs, and the fixed score distribution is therefore asymmetric. Half of the items receive scores of 0 (*not applicable to the patient*), and progressively fewer items receive higher values. We chose this asymmetric distribution because (a) we are measuring primarily abnormal personality characteristics that by definition are not present in most people, (b) such an asymmetric distribution emerges naturally with most psychopathology measures (i.e., most people do not have a given form of pathology, and progressively fewer have the pathology in more extreme form), and (c) the distribution approximates the distribution generated naturally by most clinicians when they are permitted to rate SWAP items without a fixed distribution.

A guiding principle in developing the instrument was to avoid the inevitable ambiguities of meaning that arise when the same item is expected to do double duty by representing opposite ends of a seemingly bipolar dimension (e.g., is the opposite of depression the absence of depression, happiness, or mania?). Therefore, SWAP items are never "negatively" descriptive of a patient. They are descriptive to a greater or lesser extent or else they are irrelevant to describing the patient. The meaning of a score of 0 is therefore never ambiguous: The proper interpretation of the 0 category is "irrelevant to describing this patient's personality." (This is true statistically as well as conceptually, because items with lower scores have little impact on the correlation of one SWAP score profile with another profile, and items with progressively higher scores have progressively greater influence on the

[2] This can be demonstrated conceptually (see Block, 1961/1978) as well as empirically. Empirically, correlations between SWAP profiles for unrelated individuals (i.e., who do not share common diagnostic features) tend toward 0 and in many cases are negative.

correlation coefficient.[3]) There is no loss of information (about how "not descriptive" an item is) associated with this approach: For personality constructs that have an "opposite" meaning (or multiple opposite meanings), the opposite meanings are captured by separate items. For example, the SWAP–II includes the items "Tends to be conscientious and responsible" and "Tends to be unreliable and irresponsible (e.g., may fail to meet work obligations or honor financial commitments)," thereby capturing both poles of the responsibility construct; likewise for other items where opposite meanings may apply.

The fixed score distribution does not, as some have suggested, artifactually attenuate comorbidity between related or overlapping PD diagnoses. Categories 4 through 7 of the fixed distribution include 44 items, giving assessors ample room to include multiple forms of pathology among the items designated as descriptive of the patient. Indeed, the distribution would allow an assessor to include enough criteria from each Axis II disorder to meet *DSM–IV* diagnostic cutoffs for every PD. Empirically, comorbidities among the *DSM–IV* PDs assessed via the SWAP are comparable to those assessed using structured interviews. However, they are substantially lower when assessing diagnostic groupings derived empirically (described in a later section), which "carve nature at the joints" better than the current *DSM–IV* categories (much as dimensions derived via factor analysis of self-report instruments show lower intercorrelations than *DSM–IV* diagnoses). In this case, reduced comorbidity is a function of a better taxonomy, not an artifact of a fixed score distribution.

Developing an Appropriate Item Set

The SWAP–II item set was developed and refined using standard psychometric methods. The biggest difference between the SWAP and the major self-report instruments used in PD research (e.g., the Schedule for Nonadaptive and Adaptive Personality [SNAP; Clark, 1993] and the Dimensional Assessment of Personality Pathology [DAPP, Livesley & Jackson, in press]) is that it is designed for use by clinician-informants and samples more explicitly from the universe of constructs considered important by clinicians who treat personality pathology (including not only Axis II constructs but also, for example, ways of regulating emotions, capacity for intimate relationships, characteristic motives, coping strategies, perceptions of self and others, etc.). The adult item set has undergone two major revisions since first published; the SWAP–II is the third edition of the instrument. The adolescent item set has undergone one major revision; the SWAP–II–A is the second edition of the instrument.

The SWAP–II item sets incorporate constructs from a wide range of sources including PD diagnostic criteria in *DSM–III* through *DSM–IV* as well as constructs described in the text and appendixes; selected Axis I items that could reflect personality processes (e.g., depression and anxiety); a survey of the clinical literature on PDs written over the past 50 years; research on coping and defensive processes; research on interpersonal pathology in PD patients; research on normal personality traits; research on the psychological characteristics of PDs conducted since the development of Axis II; and observations derived from pilot clinical interviews (for a more detailed description of sources, see Westen & Shedler, 1999a). The item set for the SWAP–II–A also includes constructs drawn from research on adolescent development, personality, and psychopathology (Westen & Chang, 2000).

Items are written in a manner close to the data (e.g., "Tends to get into power struggles," or "Is capable of sustaining meaningful relationships characterized by genuine intimacy and caring"); statements that require inference about internal processes are written in straightforward, jargon-free language (e.g., "Tends to see own unacceptable feelings or impulses in other people instead of in him/herself"). Writing items in this jargon-free manner minimizes unreliable interpretive leaps and makes the item set useful to all clinicians regardless of their theoretical orientation. Note that the items are also written in the form of diagnostic criteria. Items that prove to be empirically diagnostic for a disorder can therefore be used directly as candidate diagnostic criteria without the need for translation from the language of self-report to the language of clinical description, with the possibility for "slippage" of meaning that such translations entail. Many items capture personality styles and problems that are not severe enough to warrant an Axis II diagnosis, and many cover domains of healthy functioning, allowing the instrument to provide a comprehensive assessment encompassing personality strengths as well as pathology.

The earlier SWAP–200 item set was the product of a 7-year iterative item revision process incorporating the feedback of hundreds of clinician-consultants who described their patients using earlier drafts of the instrument (Shedler & Westen, 1998). We asked the clinician-consultants whether they were able to describe everything they considered psychologically important about their patients and obtained feedback about perceived omissions as well as item wording and clarity. We added, rewrote, and revised items based on this feedback, then asked new clinician-consultants to describe new patients. We repeated this process over many iterations over a period of 7 years. The current SWAP–II reflects the additional input of approximately 2,000 clinicians of all major theoretical orientations. With each revision of the item set, we also performed item analyses to identify empirically redundant items (i.e., that were excessively correlated), items that showed little power to discriminate (e.g., little variance), and so on.

We formally evaluated the comprehensiveness and content validity of the resulting item set. In the two large-sample studies described in this article, we asked clinicians who used the SWAP–II (N = 1,201) and SWAP–II–A (N = 950) to rate the following statement (5-point scale; 1 = *strongly disagree*, 5 = *strongly agree*): "The SWAP allowed me to express the things I consider important about my patient's personality." Eighty-four percent and 86% of clinicians, respectively, agreed or strongly agreed (less than 5% disagreed). The results did not differ by profession (psychiatry or psychology) or clinician theoretical orientation, suggesting that clinicians of all theoretical orientations found the item set equally relevant and useful. We are unaware of any other personality item set that has been evaluated in this manner for clinical comprehensiveness.

[3] One can readily verify this computationally. Items with low scores have the least deviation from the mean SWAP item score and therefore have the least impact on the numerator in the formula for computing Pearson's *r*.

Why Use Clinicians as Informants?

Self-reports of PD characteristics (whether obtained by questionnaire, or obtained through structured interviews that rely on overt responses to direct questions) often show only modest correlations with aggregated informant reports or consensus diagnoses using all available data, ranging from .00 to .60, with median cross-informant correlations ranging from .20 to .40 (Clifton, Turkheimer, & Oltmanns, 2005; Klein, 2003; Klonsky, Oltmanns, & Turkheimer, 2002; Pilkonis, Heape, Ruddy, & Serrao, 1991). Increasingly, research has found that informant reports predict incremental variance in relevant criterion variables (e.g., adaptive functioning in relationships, work, etc.) both concurrently and longitudinally, even after controlling for self-report. It appears, then, that others are able to recognize aspects of personality that people tend not to acknowledge in themselves. Conversely, the same studies have shown that individuals with personality pathology know some things about their inner states that they may not disclose to acquaintances.

Self-report data tend to be more predictive of internalizing pathology (which people may not share with friends or acquaintances), whereas observer reports tend to be more predictive of externalizing pathology (which people may not share with themselves; see Fiedler, Oltmanns, & Turkheimer, 2004). Clinically trained observers have advantages over laypeople in assessing both of these domains: If a patient has come to them for help, they generally have access to both internal distress and to less socially desirable aspects of the patient's personality that are revealed through interpersonal interaction and descriptions of interactions.

A second reason for using clinical informants is that much human behavior reflects consciously unreportable (implicit) as well as reportable (explicit) psychological processes. This applies to personality processes just as it applies to other areas of psychological functioning (Westen, 1998; Wilson, Lindsey, & Schooler, 2000), and some PD investigators have turned to implicit measures borrowed from cognitive science for this reason (Korfine & Hooley, 2000). Many personality processes may be inaccessible via self-report by virtue of cognitive architecture (i.e., people do not have access to them whether they would like to or not), whereas others may be inaccessible due to denial, self-deception, or self-presentation.

Shedler, Mayman, and Manis (1993) demonstrated that widely used self-report measures (e.g., of neuroticism) could not distinguish between psychologically healthy individuals and psychologically troubled individuals who maintained a façade of mental health based on defensive denial (termed *illusory mental health*). Moreover, illusory mental health was associated with a pattern of physiological reactivity linked to heart disease and other physical illnesses. Although self-report instruments could not distinguish genuine from illusory mental health, clinical assessors could do so using written autobiographical narrative material (in one study) and unstructured clinical interviews (in a second study that replicated the findings in an independent sample). Considerable research suggests many personality processes that are not readily assessed via self-report can be reliably and validly assessed based on narrative material (Cousineau & Shedler, 2006; Dozier & Kobak, 1992; Main, Kaplan, & Cassidy, 1985). In fact, this is the approach taken by clinicians of all theoretical orientations when assessing personality pathology (for empirical evidence, see Westen, 1997).

The logic is analogous to the logic of intelligence testing. Psychologists do not typically measure IQ with items such as "I am good at manipulating images in my mind" or "I know a lot of big words." Responses to such questions would likely have some validity and generate statistically significant correlations with relevant criterion variables. Nevertheless, this is not the optimal way to assess intelligence. Instead of asking people their *opinions* about their vocabulary skills, psychologists present vocabulary words and draw independent conclusions regarding performances. Likewise, skilled clinicians elicit narrative information relevant to assessing personality processes (such as ways of regulating emotions, capacity for intimate relationships, characteristic motives) and draw independent conclusions. The extent to which measures derived from these clinical observations are reliable and valid is an empirical question and the subject of much of the rest of this article.

Since the publication of Meehl's (1954) classic book, *Clinical vs. Statistical Prediction*, there has been a widely held belief in psychology that clinicians cannot make reliable or valid observations. This is a misreading of Meehl's book as well as the findings of subsequent research on clinical versus statistical prediction (see, e.g., the important meta-analysis by Grove, Zald, Lebow, Snitz, & Nelson, 2000; for a detailed discussion of this topic, see Westen & Weinberger, 2004). The misunderstanding arises because the term *clinical* has been used to mean two different things: *Clinical* as used by Meehl refers to subjective, informal, nonquantitative ways of *combining data* to generate predictions without the use of statistical methods. *Clinical* as used colloquially refers to any information provided by clinical practitioners.

Meehl demonstrated that statistical prediction is superior to prediction made without the benefit of quantified data and statistical methods. He never suggested that clinicians cannot provide accurate information about their patients. Research in cognitive and perceptual psychology has shown that accrual of experience leads to more differentiated concepts and greater capacity for accurate discrimination in every perceptual domain ever studied. There is no reason to believe that this does not apply to clinical psychologists and psychiatrists, just as it applies to physicians, architects, chess players, and anyone else who gains high levels of exposure to specialized information. Indeed, if years of clinical training and practice in psychology or psychiatry confer no added expertise in identifying pathology, this would make clinical practice in the mental health field unique among all domains of human experience.

The method we employ using the SWAP is, in fact, what Meehl (1954) would have termed *statistical prediction*, not clinical prediction. Specifically, it is statistical prediction using quantified clinician data as input. The approach relies on clinicians to do what they can do well, namely, making specific observations and inferences about individual patients they treat and know well or interview systematically. It relies on statistical algorithms to do what they do well, namely, aggregating data to derive reliable, valid scales and indices and predict relevant criterion variables (cf. Sawyer, 1966). Note that the SWAP method does not ask clinicians to *predict* anything. It asks them only to describe what they observe in a systematic and quantifiable manner.

Reliability and Validity

Numerous studies have demonstrated the reliability and validity of SWAP data, even in field trials with clinicians using the instrument for the first time. (In our own laboratory, where we score the SWAP based on systematic clinical interviews, we provide assessors with training and practice, just as with other PD interviews and measures.) The best current evidence suggests that the well-documented unreliability of clinical diagnoses is an artifact of failure to use appropriate psychometric instruments to quantify clinicians' observations and inferences. We summarized the data on this broader issue elsewhere (Westen & Weinberger, 2004) but briefly describe some pertinent findings here.

In several studies, both the adult and adolescent versions of the SWAP predicted a range of relevant external criteria, from those that are relatively objective to those that require greater inference. These include, for example, history of suicide attempts and psychiatric hospitalizations; adaptive functioning assessed by measures such as the Global Assessment of Functioning index (GAF); family history variables such as psychosis in first- and second-degree relatives; and developmental variables, including being raised by a substance-abusing parent or guardian, childhood history of physical abuse, childhood history of sexual abuse, and problems with parental bonding and attachment (Bradley, Jenei, & Westen, 2005; Russ, Heim, & Westen, 2003; Shedler & Westen, 2004a).

A methodological limitation of early SWAP studies is that the same informant (the treating clinician) provided both SWAP data and data on criterion variables. However, many of the external criterion variables concern objective matters that require no inference or interpretation (e.g., history of suicide attempts, history of psychiatric hospitalizations). Also, a number of studies have correlated SWAP data with data from independent informants with equally impressive results, suggesting that validity findings cannot be explained as an artifact of shared method variance.

In one study (Westen & Muderrisoglu, 2003), two clinicians independently described a sample of outpatients using the SWAP–200 after conducting (or observing on videotape) the CDI, a systematic clinical interview (approximately 2½ hr in length) designed to resemble but systematize interview procedures used by experienced clinicians of all theoretical orientations (Westen, 1997). The investigators assessed interrater reliability for each of the *DSM–IV* Axis II PDs as well as for a set of PD diagnoses derived empirically in prior research (Westen & Shedler, 1999b). Median interrater reliability for all diagnostic scales exceeded .80. To assess validity, the investigators correlated SWAP–200 diagnostic scores obtained via the interviews with diagnostic scores provided by the treating clinicians based on longitudinal clinical observation of the patient over extended time periods. Median validity coefficients were again in the range of .80, with discriminant validity coefficients small to moderate for *DSM–IV* diagnoses and hovering near 0 for the empirically derived diagnoses. A follow-up study used the same data to examine the reliability and validity of trait scores derived from the SWAP–200 via factor analysis (Westen & Muderrisoglu, 2006). Interrater reliability by interview was once again high, with median correlations between independent interviewers of .82. Convergent and discriminant validity (assessed by cross-informant agreement) were also strong, with a median convergent validity coefficient of .66 and a desirably low median discriminant validity coefficient of –.06.

In a third study, conducted by an independent research team (Marin-Avellan, McGauley, Campbell, & Fonagy, 2005b), investigators applied the SWAP–200 to audiotaped Adult Attachment Interviews (Main et al., 1985) plus chart records for a sample of inpatients at a maximum security forensic hospital. Interrater reliability between independent assessors was high for all SWAP–200 PD scales, with a median interrater correlation of .91. The SWAP–200 proved superior to the Structured Clinical Interview II for *DSM–IV* Axis II Personality Disorders (SCID–II; First et al., 1995) in predicting aggressive ward behavior, independently assessed by ward nurses (blind to other data) using a 49-item interpersonal circumplex rating scale. SWAP antisocial PD scores correlated significantly with dominance behavior and coercive behavior on the ward and correlated negatively with submissive and compliant ward behavior. These findings were preliminary, based on a small sample ($N = 30$). In a subsequent report based on a larger sample ($N = 60$; Marin-Avellan, McGauley, Campbell, & Fonagy, 2005a), SWAP–200 PD scores remained superior to SCID–II diagnoses in predicting ward behavior and were also predictive (unlike the SCID–II) of patients' index offense, notably whether it was violent or nonviolent.

In the most definitive study to date (Westen, Waller, Blagov, Shedler, & Bradley, 2007), we collected SWAP data by interview using the CDI and assessed criterion variables by self-report and interviewer report from independent assessors blind to the CDI and SWAP. The sample was an inner-city, primarily African American sample of 150 primary care patients. Validity of SWAP–II factors (described below) from our normative sample appeared to be as high or nearly as high for cross-informant correlations as for the single-informant correlations we reported in our earlier studies. For example, the SWAP–II psychopathy factor correlated .40 with self-reported arrest history (scored categorically) and .57 with SNAP antisocial PD. SWAP emotional dysregulation correlated .36 with self-reported suicide attempts, .49 with self-reported emotional dysregulation, and .53 with SNAP borderline PD. Both SWAP scales correlated negatively with self-reported and interviewer-rated global functioning. This was the first large-sample study of the validity of SWAP–II factors using a true multitrait–multimethod matrix.

Taxonomizing Personality Pathology

The approach to revising Axis II we have been pursuing over the past decade is one that might be described as a virtual field trial, which uses large practice networks of doctoral-level clinicians to provide quantified data on a patient in their care, and applies a range of statistical procedures to aggregate the data for taxonomic purposes. Like research using the DAPP and the SNAP, we utilize what is essentially a construct validation procedure (Livesley & Jackson, 1992; Millon, 1991). We begin with a broad item set (200 items) developed through standard content validation procedures; use statistical procedures to derive diagnostic scales empirically; and then test the derived diagnostic scores for characteristics internal to the diagnostic system (e.g., reliability, diagnostic overlap) and external to the diagnostic system (i.e., validity, examining correlations with criteria not used to define the diagnoses), as well as for clinical utility.

We recently completed two large projects on the classification and diagnosis of personality pathology using the approach described here, one focusing on adults and the other on adolescents. Because these two large normative studies used similar methodology, we describe their shared methods and note any differences here. We then describe what can be done with these data and some of the findings to date, focusing, for parsimony, primarily on the adult data. The research is described in greater detail elsewhere (Russ, Bradley, Shedler, & Westen, in press; Westen, Nakash, Thomas, & Bradley, 2006; Westen, Shedler, & Bradley, 2006).

We contacted a random national sample of psychiatrists and psychologists from the membership registers of the American Psychiatric Association and the American Psychological Association. More than one third of clinicians agreed to participate for a consulting fee of $200. We asked clinicians to describe "an adult [adolescent] patient you are currently treating or evaluating who has enduring patterns of thoughts, feeling, motivation or behavior—that is, personality problems—that cause distress or dysfunction." To obtain a broad range of personality pathology, we emphasized that patients should have problematic personality characteristics but need not meet criteria for a PD diagnosis. Clinicians reported that these broad criteria applied to most of their patients, suggesting that we obtained a sample representative of most patients presenting for clinical treatment. For adolescents, we obtained a stratified random sample, stratifying on age and gender.

To avoid selection biases, we directed clinicians to consult their calendars and select the last patient they had seen during the previous week who met study criteria, and we conducted formal checks to verify that the clinicians followed these procedures. Both samples were diverse in terms of level of patients' pathology, ethnicity, age, socioeconomic status, site at which they were treated, and so forth, as well as clinicians' theoretical orientation and professional degree.

The core battery of measures required approximately 2 hr to complete. Aside from the SWAP, we collected a range of other measures to assess demographics, diagnostic information, and potential criterion variables for assessing validity of SWAP-derived diagnoses and indices. These included measures of adaptive functioning, Axis I and Axis II pathology, family history of psychopathology in biological relatives, developmental history (e.g., disrupted attachments, poverty, physical and sexual abuse, parental criminality), and treatment response (assessed concurrently in adults and prospectively in adolescents). We also examined test–retest reliability (or temporal stability) of SWAP–II factor scores by contacting a subset of clinicians after a time interval and asking them to assess their patients again. The median test–retest reliability ($N = 94$) for SWAP–II factor scores was .85 after an interval of 4 to 6 months. (Westen et al., 2007).

To obtain Axis II diagnoses independent of SWAP data, we included two additional measures. The first was an Axis II checklist, which listed all Axis II criteria from *DSM–IV* for all disorders, randomly ordered, which allowed us to make both dimensional and categorical Axis II diagnoses based on *DSM–IV* algorithms. Data from this checklist mirror data from structured PD interviews, with similar patterns of comorbidity and similar associations with measures of adaptive functioning.

The second method for obtaining Axis II diagnoses independent of SWAP data relied on what we termed PD *construct ratings*. We asked clinicians to rate the extent to which the patient resembled or "matched" each *DSM–IV* PD construct, irrespective of specific diagnostic criteria (5-point scale; 1 = *little or no match*, 5 = *very good match, prototypical case*). To guide the clinicians, we reproduced the single-sentence summary that introduces each disorder in *DSM–IV* (e.g., "The essential feature of Borderline Personality Disorder is a pervasive pattern of instability of interpersonal relationships, self-image, and affects, and marked impulsivity that begins by early adulthood and is present in a variety of contexts"; American Psychiatric Association, 1994, p. 650). Additional scale anchors indicated that ratings ≥4 signified "caseness," meaning the clinician believed the patient "had" the PD. We included construct ratings because one goal of the study was to refine the current Axis II diagnoses, and we wanted to avoid circularity that could arise by diagnosing patients by a particular set of criteria and then testing whether those same criteria best define the disorder.

For taxonomic purposes, these data can be used in four primary ways: (a) to refine the diagnostic constructs and criteria for PDs currently included in the *DSM*, (b) to identify diagnostic groupings empirically, without presupposing the *DSM–IV* classification, (c) to identify the trait structure of personality pathology using an item set designed for clinically sophisticated informants, and (d) to identify diagnostic subtypes and lower order constructs (latent traits) underlying diagnostic syndromes. We describe each of these uses in turn after a brief clarification regarding the meaning of "dimensional" diagnosis.

Meanings of Dimensional Diagnosis

An important consideration for personality researchers is whether diagnostic assessment should focus on personality syndromes or personality traits. *Syndromes* are multifaceted constellations of personality processes (encompassing cognition, affectivity, interpersonal functioning, impulse regulation, etc.; American Psychiatric Association, 2000, p. 686) that are understood to be interdependent. All editions of the *DSM* to date have focused on syndromes. In contrast, *trait* approaches focus on discrete dispositions typically derived through factor analysis (e.g., extroversion, neuroticism).

Some investigators mistakenly conflate trait approaches with dimensional diagnosis, and syndromal approaches with categorical diagnosis (e.g., the *DSM–IV* typology of PDs is syndromal and also categorical). However, these are independent considerations whose association is purely historical (Westen, Gabbard, et al., 2006). The dimensional/categorical distinction refers to whether people are assumed to fall into discrete categories or to vary along a continuum. The syndromal/trait distinction refers to whether the unit of diagnosis is a constellation of interrelated personality characteristics or separate characteristics.

In the following sections, we describe two syndromal approaches and one trait approach, all based on SWAP–II data. All of the approaches are dimensional. In the case of the syndromal approaches, diagnostic groupings are defined by empirically derived prototypes—descriptions that represent each diagnostic syndrome in its ideal or pure form (based on all 200 SWAP items). Individual patients are diagnosed dimensionally (on a continuum) based on the degree of resemblance or match with the prototype (Westen, Shedler, et al., 2006).

Revising PD Constructs and Criterion Sets

The first approach, and the one most continuous with past efforts to revise *DSM* criterion sets, is to identify candidate diagnostic criteria that may be more useful than the current criteria. This approach is similar to one we undertook in our prior normative study of the SWAP–200 ($N = 530$). However, the present methodology addresses some limitations of the earlier project, of which three are most important. First, in our prior normative adult study, we asked clinicians to select a patient with a particular *DSM–IV* PD, which could have biased the findings in the direction of reproducing *DSM–IV* constructs and diagnostic criteria. Second, in our earliest studies clinicians were allowed some leeway in selecting patients (if they had more than one patient in their practice with the applicable diagnosis), introducing potential sampling bias. Finally, the composite descriptions of disorders (e.g., borderline PD) in our prior studies identified the most *characteristic* features of each disorder (i.e., the items most characteristic of the average patient with the disorder) but did not necessarily identify the most *distinctive* features of each disorder (i.e., those that would distinguish it from near-neighbor disorders). Negative affectivity, for example, appears to be a central component of borderline PD, but it does not distinguish patients with borderline PD from those with other disorders, such as dependent and avoidant PD.

In the present studies, we did not ask clinicians to describe a patient with a specific PD but instead used a procedure to sample patients randomly from the clinicians' practices. To identify patients with each *DSM–IV* disorder (whose SWAP data could be aggregated to identify the most diagnostic items or criteria), we used two methods: (a) selecting patients who met *DSM–IV* diagnostic criteria for each disorder using current criteria and cutoffs, based on the Axis II Checklist; and (b) selecting patients who strongly resembled or "fit" each *DSM–IV* PD diagnosis based on PD construct ratings (for which ratings of 4 and 5 were defined as caseness). This latter method uses the categories in *DSM–IV* without presuming the specific criteria specified by the manual.

To identify the most characteristic as well as the most distinctive features of each disorder, we generated two composite descriptions for each disorder. The first aggregated the raw SWAP item scores across all patients diagnosed with the disorder, producing a composite description of the average or typical patient with the disorder. We created this description simply by sorting the SWAP items from highest to lowest with respect to their ranking in the composite description, then examining the top 18 to 30 items (i.e., the items that fell into the two or three most descriptive piles according to the fixed distribution of the Q sort). The second composite was created by aggregating SWAP item scores of patients meeting criteria for a disorder after standardizing (z-scoring) the items across patients (i.e., transforming the distribution of items so that the mean was 0 and the standard deviation was 1). This latter procedure deemphasizes items that have high means for most patients (e.g., dysphoric affect, which is common in a psychiatric sample) and weights items more heavily that uniquely distinguish patients diagnosed with a given disorder within the sample. Thus, we employed two alternate methods of identifying patients with each disorder and two alternate methods of creating aggregate or composite descriptions of these patients. Figure 1 depicts this approach schematically.

The data on borderline PD are instructive (Bradley, Shedler, & Westen, 2006). Examination of the highest-ranked items in these two composites identified a stable core of 15 items that are both characteristic and distinctive of BPD (e.g., "Emotions tend to spiral out of control, leading to extremes of anxiety, sadness, rage, etc."), a stable core of items that are characteristic but not distinctive (primarily indicators of negative affectivity; e.g., "Tends to feel unhappy, depressed, or despondent"), and a set of features that emerge only under extreme stress that are highly distinctive of BPD but not stable (e.g., "Tends to engage in self-mutilating behavior"). These last features appear to account for much of the apparent instability of the diagnosis over time in longitudinal research.

The findings show a striking convergence with recent data from longitudinal studies distinguishing (a) a stable core of BPD features that includes both emotional dysregulation and negative affectivity and leads to rank-order stability of patients over time; and (b) an unstable set of behavioral indicators that wax and wane or diminish over time, particularly in samples in which many patients appear to be in long-term, stable treatment relationships (e.g., McGlashan et al., 2005; Zanarini, Frankenburg, Hennen, & Silk, 2003).

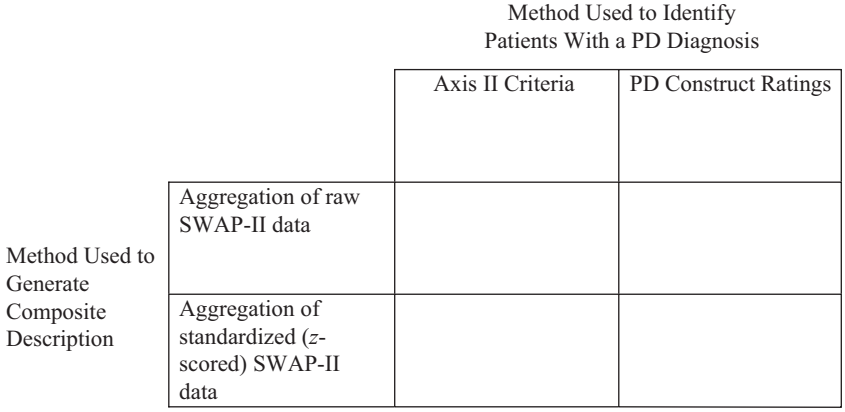

Figure 1. Procedure for generating composites of patients with a given personality disorder (PD). SWAP–II = Shedler-Westen Assessment Procedure II.

Deriving Personality Configurations Empirically

The data analytic approach described previously presumes the basic accuracy of the *DSM–IV* taxonomy of PDs and attempts to optimize diagnostic criteria within the framework of that taxonomy. A second approach is to use procedures such as Q factor analysis and latent class analysis to identify diagnostic groupings empirically. We do not extensively review our prior research using the SWAP–200 and SWAP–200–A, which has been described elsewhere, but address two issues here. First, we do not believe we should assume a priori that personality falls into discrete types, and therefore prefer procedures that allow us to determine the latent dimensions that underlie the data, whether they be discrete trait dimensions or syndromes (complex constellations of interrelated personality processes). Second, although we used Q factor analysis extensively in our prior research using the SWAP–200, nothing in the structure of our data weds us to any particular data analytic procedures. Procedures such as the novel dovetailing of latent trait and latent class analysis recently described by Krueger, Markon, Patrick, and Iacono (2005) provide alternative methods we are exploring.

Our preference for Q factor analysis thus far stems from two basic considerations. First, empirically, it consistently provides the cleanest groupings of patients we have seen in the literature over the past 30 years of research employing a range of person-centered data analytic strategies to multiple psychiatric disorders. The Q-factors that emerged from our studies of both adult and adolescent patients using the SWAP–200 and SWAP–200–A are clear, clinically recognizable, theoretically sensible, and have predictable external correlates, whether we Q-factor a sample of adult patients selected for having any PD (Westen & Shedler, 1999b), adolescent patients with any form of personality pathology (Westen et al., 2003), patients with borderline PD (Bradley, Zittel, & Westen, 2005), or those with eating disorders (Westen & Harnden-Fischer, 2001). Where tested, these diagnostic groupings have virtually all replicated across samples. The coherence and replicability of these prototypes stand in contrast to most of the diagnostic groupings that have emerged in three decades of cluster analytic research as well as recent research applying latent class analysis to a range of disorders.

The second reason for our preference thus far for Q factor analysis is its mathematical elegance and well-known mathematical properties. Q factor analysis is simply conventional factor analysis with the rows and columns of data transposed. Thus, everything one knows about factor analysis applies to Q factor analysis. With three exceptions, any criticism of Q factor analysis is simultaneously a criticism of the factor analytic procedures that have provided the foundation for much of the dimensional measurement in psychology and psychiatry (other than measurement of the *DSM* categories). The first exception is that, because variables and cases are transposed in Q factor analysis, with large data sets the number of variables (patients) will exceed the number of cases (items). Pragmatically, the primary limitation is in the estimation (factor extraction) procedures that can be used. Second, if the goal is to derive discrete, mutually exclusive classes, Q factor analysis is not the optimal method because, like factor analysis, it maximizes the variance accounted for by latent variables across diagnoses, rather than forcing patients into mutually exclusive groups. However, we do not assume that personality syndromes fall into mutually exclusive classes. Taxonicity is an empirical question. Third, in the absence of a fixed distribution and adequate item coverage (e.g., including items indicative of psychological health, which allow the assessor to distinguish, for example, a patient with narcissistic PD from a much higher functioning patient who has narcissistic personality features but not a PD), two patients with opposite profiles, or with the same profile at very different levels (e.g., severe and mild), could in principle load on the same Q-factor (Waller & Meehl, 1998). However, with a fixed distribution, two patients at different levels of the same disorder cannot, mathematically, have similar profiles, and empirically, one rarely obtains Q factors with mixed (positive and negative) loadings (which would indicate patients with opposite pathology loading on the same Q-factor).

We describe here some preliminary findings from ongoing data analyses with our large normative samples of adults and adolescents. In a first set of analyses, we used Q factor analysis to identify naturally occurring diagnostic groupings. We uncovered a hierarchical structure in both the adult and adolescent data sets (see Figure 2 for the structure that emerged in the adult sample). At the superordinate level for both adults and adolescents are three broad diagnostic groupings or clusters: internalizing, externalizing, and borderline. (We also obtained a high-functioning spectrum in each data set, which includes patients with personality syndromes such as obsessive-compulsive, which has often predicted good rather than poor adaptive functioning.) What is striking about this super-

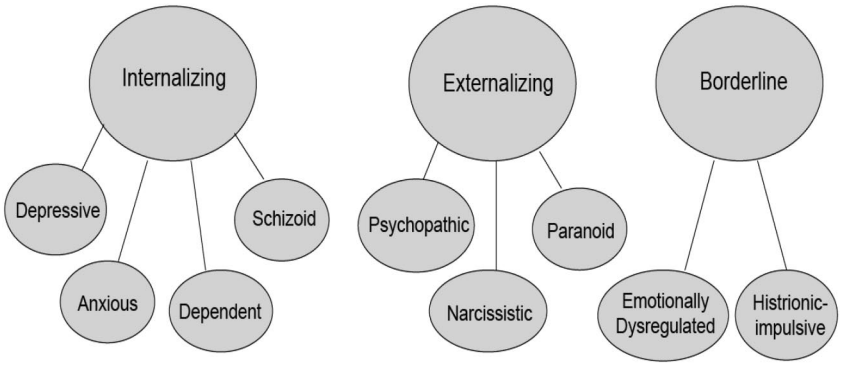

Figure 2. Hierarchical structure of empirically derived personality constellations.

ordinate structure is that we reproduced, using person-centered analysis of personality syndromes, what several researchers have identified using variable-centered analyses with primarily Axis I disorders, namely internalizing and externalizing spectrum dimensions (Brown, Chorpita, & Barlow, 1998; Krueger, 1999; Krueger et al., 2002; Watson & Clark, 1992). Additionally, we identified a borderline spectrum that contains elements characteristic of both internalizing and externalizing pathology. The findings suggest the possibility of integrating our understanding of many Axis I disorders with our understanding of personality and personality pathology.

By conducting a second-order Q factor analysis, in which we Q-factored the profiles of patients who loaded >.40 on any of the three superordinate factors, we arrived at the structure in Figure 2. As can be seen, we identified a number of nonredundant diagnostic groupings, many of which resemble Axis II diagnoses and some of which do not. The groupings, we believe, make better sense than the *DSM–IV* clusters, which have never had strong empirical support.

To give an example of these second-order diagnoses, we reproduce here some of the items most descriptive of the psychopathic PD diagnosis, in descending rank order: "Takes advantage of others; has little investment in moral values," "Has little empathy; seems unable or unwilling to understand or respond to others' needs or feelings," "Experiences little or no remorse for harm or injury caused to others," "Tends to blame own failures or shortcomings on other people or circumstances," "Has little psychological insight into own motives, behavior, etc.," "Tends to act impulsively (e.g., acts without forethought or concern for consequences)," "Tends to be deceitful; tends to lie or mislead," "Is prone to intense anger, out of proportion to the situation at hand," "Tends to show reckless disregard for the rights, property, or safety of others," "Tends to abuse drugs or alcohol," and "Tends to engage in unlawful or criminal behavior." What is striking is that this relatively pure diagnosis, which strongly resembles the psychopathy construct, emerged despite the fact that very few of the patients with *DSM–IV*-diagnosed antisocial PD were noncomorbid; thus, what Q factor analysis identified was a clinically and empirically coherent signal amidst considerable noise. It is noteworthy that several characteristics ranked as most descriptive of this diagnosis are not included among the current *DSM–IV* criteria for the disorder (e.g., lack of empathy, which had the second highest ranking out of 200 items).

Identifying Trait Dimensions

The third way of using the data is similar to efforts to devise trait approaches to personality classification using patient self-reports, except that we apply factor analysis to an item set designed for experienced clinical observers. In our prior studies, we subjected the SWAP–200 and SWAP–200–A to factor analysis and identified several factors that resemble the "Big Four" psychopathology factors, such as a hostility factor that resembles low agreeableness, as well as factors that are not represented in four-factor space at all (e.g., sexual conflict, schizotypy, or thought disorder). We also identified some potentially useful diagnostic distinctions, such as the distinction between negative affectivity and emotional dysregulation, which increasingly appear to be distinct constructs, as reflected in the difference between stable dysthymia and borderline PD. Both of these disorders are characterized by high scores on neuroticism or negative affectivity, but only the latter is characterized by emotional dysregulation (Miller & Pilkonis, 2006; Shedler & Westen, 2004a; Westen et al., 2005).

In our most recent studies, undertaken in collaboration with Niels Waller, we derived the trait structure of the SWAP–II using tetrachoric correlations to minimize the impact of differential item skewness on factor structures (Westen et al., 2007). A 17-factor solution provided the most clinically and theoretically coherent solution, and it resembled in many respects the factor structure of the SWAP–200. The factors (and highest loading items) included the following (not all listed here): Psychological Health, Psychopathy (e.g., "Tends to be deceitful," "Tends to take advantage of others"), Hostility (e.g., "Tends to be critical of others," "Tends to be angry or hostile"), Narcissism (e.g., "Has fantasies of unlimited success, power, beauty, talent, brilliance, etc.," "Has an exaggerated sense of self-importance"), Obsessionality (e.g., "Tends to be overly concerned with rules, procedures, order, organization, schedules, etc.," "Tends to become absorbed in details, often to the point that s/he misses what is significant"), Depression (e.g., "Tends to feel life has no meaning"), Emotional Dysregulation (e.g., "Emotions tend to spiral out of control"), Schizotypy (e.g., "Reasoning processes or perceptual experiences seem odd and idiosyncratic"), and Emotional Avoidance (e.g. "Is invested in seeing and portraying self as emotionally strong, untroubled, and emotionally in control despite clear evidence of underlying insecurity, anxiety, or distress"). Some of these factors are clearly interpretable from a four- or five-factor perspective, whereas others are not. For example, Obsessionality includes items related to cognitive style, not just to high conscientiousness; Emotional Dysregulation correlates only moderately with negative affectivity; and Schizotypy is not represented in four-factor space at all.

Identifying Latent Traits and Subtypes

Although these three approaches (refining the current PD taxonomy, empirically identifying naturally occurring diagnostic groupings, and identifying trait dimensions via factor analysis) represent the most comprehensive ways the SWAP can be used for taxonomic purposes, the methodology can also be used to address more specific goals. For example, it can be used to clarify the lower order constructs (latent traits) composing current diagnoses, such as borderline PD. We recently analyzed the data on patients meeting borderline PD criteria in the adult data set and identified six factors: negative affectivity, hostility, emotional dysregulation, unstable identity/impulsivity, attachment dysregulation, and self-harm (Bradley, Shedler, & Westen, 2006).

Similarly, using person-centered (syndromal) analyses, researchers can identify subtypes of current PD diagnoses. For example, using the adult sample, we applied Q factor analysis to all patients meeting independent criteria for narcissistic PD (Russ, Bradley, Shedler, & Westen, in press). We identified three subtypes: Grandiose/Malignant, Fragile, and High Functioning. Fragile narcissists were characterized by both feelings of grandiosity and feelings of inadequacy, suggesting that the former is a defense against the latter. Grandiose/Malignant narcissists showed none of the underlying vulnerability and appeared closer to a psychopathic spectrum disorder. High functioning narcissists were self-absorbed

and attention seeking, but were capable of stable attachments and were generally able to function effectively in the world.

In another study, we applied Q factor analyses to the SWAP–II–A data of the 138 adolescents in the adolescent data set who met *DSM–IV* criteria for attention-deficit/hyperactivity disorder (Levin, Bradley, & Westen, 2006). Q factor analysis identified four personality subgroups: Psychopathic, Socially Withdrawn, Emotionally Dysregulated, and High Functioning. The four groups differed on a range of criterion variables, including Axes I and II pathology, adaptive functioning, developmental and family history variables, and treatment response. Of particular interest, the Psychopathic subtype appears to represent a particularly malignant externalizing disorder characterized by early onset of severe conduct problems such as animal torture and interpersonal violence, family history of criminality in biological relatives, and the relative absence of anxiety disorders in biological relatives as compared with the other subtypes. The personality subtypes also showed substantial incremental validity in predicting global functioning and treatment response. Other analyses have suggested potential forensic uses of the SWAP. In one study, we identified subtypes of partner-violent men that resemble those found using other methods, notably a psychopathic subgroup and a more fragile, dependent, emotionally dysregulated subgroup that differed on a range of criterion variables (Fowler & Westen, 2007).

Limitations and Challenges

As noted earlier, a limitation of our large-sample virtual field trials regards the validity analyses, because, as in most PD research, we relied on a single informant. In our case, the single informant was the treating clinician; in most other personality research, the single informant is the patient.[4] However, it was never our intent to limit data using the SWAP to studies in which clinicians are the primary or only informants. We chose that strategy as an initial research step because it allowed us to collect very large samples that made taxonomic work possible.

We are currently engaged in three studies that will allow us to assess construct validity using the kind of multitrait–multimethod matrices that represent the gold standard described by Campbell and Fiske (1959). The first, described briefly already, is a study of several hundred largely African American nonpatients in inner-city Atlanta, on whom we are collecting personality data using both the SWAP and SNAP as well as data on a range of other psychological and etiologic variables, including Axis I pathology, adaptive functioning, resilience (in a group heavily exposed to stressors such as interpersonal violence and poverty), history of suicide and criminality, molecular genetics, and developmental history. The research is part of an ongoing study focused on genetic and environmental predictors of the development of posttraumatic stress disorder with a target sample size of 400.

A second study addresses another crucial criterion variable identified by Robins and Guze (1970) for validating an approach to classification, namely treatment response. As part of a center grant on predictors of treatment response in treatment-naïve patients with major depression, we are examining personality assessed via the SWAP (again collected using the CDI, not the treating clinician) as a predictor of response to cognitive–behavioral therapy and to two antidepressants in a projected sample of 420 patients. Other variables assessed in this study include baseline functional magnetic resonance imaging data, data on history of childhood abuse and adverse events, neuroendocrine data, and molecular genetics.

The third study, recently begun, is aimed at comparing multiple approaches to taxonomy and measurement of personality pathology in a clinical sample ($N = 240$), including *DSM–IV* diagnosis assessed by the SCID–II, a four-factor trait model assessed by self-report, and our SWAP-derived classifications. Criterion variables include etiological variables (e.g., molecular genetics and developmental adversity) and adaptive functioning assessed prospectively 18 months after initial assessment.

Future Directions: The Challenge of Scaling

The primary focus of SWAP research to date, and of this article, has been classification and taxonomy, rather than the use of SWAP instruments for clinical assessment of individual patients (for an illustration of the potential of the instrument in clinical assessment, see Lingiardi, Shedler, & Gazzillo, 2006). We have thus far not published a test manual intended for clinical users, although we are in the process of developing one. The only manual we have produced thus far was based on our earliest, preliminary studies with the SWAP–200 and was intended solely to provide basic guidance to empirical investigators who wished to collect SWAP–200 data for research purposes.

With the SWAP–II data set, we now have a representative sample from a well-defined population, and questions of population norms and scaling of diagnostic scores for purposes of clinical assessment have become more salient. There are no simple answers to the question of how best to scale a personality pathology instrument. In the past, we used *T* scores because they provided a convenient and easily computed metric. Our earlier normative sample using the SWAP–200 consisted almost entirely of PD patients (not a representative sample from a well-defined population), and the resulting *T* scores did not adjust for differential base rates of different forms of personality pathology in the population or the sample.

Using the data from our current projects, scaling is more straightforward for the current Axis II disorders, because the Axis II Checklist provides information on whether patients cross the (arbitrary) *DSM–IV* thresholds for each disorder. We are leaning toward the view that the best way to scale PD diagnoses dimensionally for the current Axis II PDs may be to use percentile scores and/or probability scores (e.g., this patient has an 83% likelihood of having narcissistic PD as defined by *DSM–IV*).

How to scale empirically derived diagnoses is a question no one has satisfactorily answered. Percentile scores and normalized *T* scores would provide readily comprehensible metrics, although neither method is without challenges. However we ultimately decide to scale empirically derived traits and syndromes, one important advantage of the SWAP approach deserves mention. Because the items are written in clinical language and describe personality *functions* (e.g., ways in which the person regulates or fails to regulate impulses, emotions, self-esteem), they can be used

[4] In our view, the patient is the primary informant regardless of whether data are collected via questionnaires or via the major structured interviews that rely on a patient's overt responses to direct questions.

to create narrative descriptions of patients in plain clinical language, allowing not only quantitative score profiles but also interpretive reports written in the language of the instrument itself (i.e., without the slippage of meaning that may occur when self-report items are translated into clinical diagnostic constructs; see Lingiardi et al., 2006; Shedler & Westen, 2007; Westen, Gabbard, et al., 2006). We are developing a computerized interpretive report organized by functional domains (e.g., emotion, emotional regulation, cognitive functioning, interpersonal functioning, experience of self) that are the targets of most forms of treatment of personality pathology.

Coda: On Ideology and Science

SWAP research has elicited polarized responses. In some quarters it has elicited strong enthusiasm; in others, it has been greeted with deep skepticism. It has also been misrepresented (e.g., Garb, 2005; Widiger & Samuel, 2005). Although any new approach may elicit skepticism from investigators committed to other paradigms (Kuhn, 1962), the extent of the polarization has surprised us. Possibly, SWAP research has inflamed ideological rifts in psychology related to the scientist-practitioner schism, which has been commented on in every American Psychological Association presidential election in recent decades, and that was acrimonious enough to divide the organization's membership and lead to the formation of the Association for Psychological Science as a separate organization in 1988. For some scientific psychologists, the limitations of clinical judgment are highly salient (Garb, 1998), discussed in terms of Barnum effects and likened to astrology (Garb & Grove, 2005).

This ideological rift is illustrated in Garb's (2005) review of clinical judgment in the prestigious *Annual Review of Clinical Psychology*. In his view, there are two traditions in psychology, which he termed *romantic* and *empiricist* (citing Wood, Nezworski, Lilienfeld, & Garb, 2003, pp. 92–94). The former is unscientific, anecdotal, and subject to bias. The latter is scientific and will ultimately prevail.

Garb cited our approach as an example of the so-called romantic tradition and stated that "using this approach, the *DSM* criteria would not be revised on the basis of research studies on the etiology, nature, and course of a mental disorder, but instead on the basis of clinicians' observations" (p. 93). This does not represent our position, and we find it surprising that someone would draw such a conclusion from our empirical publications. On the contrary, we advocate the use of clinical findings to refine diagnosis where those findings are supported by empirical evidence, demonstrate construct and criterion validity, and can be located in a coherent nomological network of empirical findings. We also reject the premise that there must be a dichotomy between clinical and empirical approaches. On the contrary, we assume that each can inform the other and that good psychology, like good science more generally, involves an interplay of observational and empirical findings. Efforts by some investigators to shut out clinical observation and deduction does not make for better science; it makes for poorer science by eliminating a rich source of potential hypotheses and by drastically limiting the range of psychological phenomena considered and investigated. All of our proposals to date have been based on empirical findings, and we advocate (and are conducting) further research that subjects our concepts to risk of disconfirmation. Criticism of the SWAP, however, has thus far relied on armchair analysis without any supporting empirical data.

Ultimately, the question of the utility of the SWAP depends not on discussions of whether particular investigators find the fixed distribution useful or problematic, or the optimal method of scaling, or debates about what clinicians can and cannot do right, but on whether the constructs and measures that emerge from it prove more or less empirically valid and clinically useful than available alternatives.

References

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

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text rev.). Washington, DC: Author.

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

Blagov, P., Bradley, R., & Westen, D. (2007). Under the Axis II radar: Clinically relevant personality constellations that escape *DSM–IV* diagnosis. *Journal of Nervous and Mental Disease, 195,* 477–483.

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. (Original work published 1961)

Bradley, R., Jenei, J., & Westen, D. (2005). Etiology of borderline personality disorder: Disentangling the contributions of intercorrelated antecedents. *Journal of Nervous and Mental Disease, 193,* 24–31.

Bradley, R., Shedler, J., & Westen, D. (2006). *Refining the borderline construct: Diagnostic criteria and trait structure.* Unpublished manuscript, Emory University.

Bradley, R., Zittel, C., & Westen, D. (2005). Borderline personality disorder in adolescence: Phenomenology and subtypes. *Journal of Child Psychology and Psychiatry, 46,* 1006–1019.

Brown, T. A., Chorpita, B. F., & Barlow, D. H. (1998). Structural relationships among dimensions of the *DSM–IV* anxiety and mood disorders and dimensions of negative affect, positive affect, and autonomic arousal. *Journal of Abnormal Psychology, 107,* 179–192.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81–105.

Clark, L. A. (1993). *SNAP (Schedule for Nonadaptive and Adaptive Personality): Manual for administration, scoring and interpretation.* Minneapolis: University of Minnesota Press.

Clifton, A., Turkheimer, E., & Oltmanns, T. F. (2005). Self- and peer perspectives on pathological personality traits and interpersonal problems. *Psychological Assessment, 17,* 123–131.

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.

Cousineau, T. M., & Shedler, J. (2006). Predicting physical health: Implicit mental health measures versus self-report scales. *Journal of Nervous and Mental Disease, 194,* 427–432.

Dozier, M., & Kobak, R. (1992). Psychophysiology in attachment interviews: Converging evidence for deactivating strategies. *Child Development, 63,* 1473–1480.

Fiedler, E., Oltmanns, T., & Turkheimer, E. (2004). Traits associated with personality disorders and adjustment to military life: Predictive validity of self and peer reports. *Military Medicine, 169,* 32–40.

First, M., Spitzer, R., Gibbon, M., Williams, J., Davies, J. B., Howes, M., et al. (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.

Fowler, K., & Westen, D. (2007). *Subtyping male perpetrators of domestic violence.* Unpublished manuscript, Emory University.

Garb, H. N. (1998). *Studying the clinician: Judgment research and psychological assessment.* Washington, DC: American Psychological Association.

Garb, H. N. (2005). Clinical judgment and decision making. *Annual Review of Clinical Psychology, 1,* 67–89.

Garb, H. N., & Grove, W. M. (2005). On the merits of clinical judgment. *American Psychologist, 60,* 658–659.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12,* 19–30.

Klein, D. N. (2003). Patients' versus informants' reports of personality disorders in predicting 7 1/2-year outcome in outpatients with depressive disorders. *Psychological Assessment, 15,* 216–222.

Klonsky, E. D., Oltmanns, T. F., & Turkheimer, E. (2002). Informant-reports of personality disorder: Relation to self-reports and future research directions. *Clinical Psychology: Science & Practice, 9,* 300–311.

Korfine, L., & Hooley, J. M. (2000). Directed forgetting of emotional stimuli in borderline personality disorder. *Journal of Abnormal Psychology, 109,* 214–221.

Krueger, R. F. (1999). The structure of common mental disorders. *Archives of General Psychiatry, 56,* 921–926.

Krueger, R. F., Hicks, B. M., Patrick, C. J., Carlson, S. R., Lacono, W. G., & McGue, M. (2002). Etiologic connections among substance dependence, antisocial behavior, and personality: Modeling the externalizing spectrum. *Journal of Abnormal Psychology, 111,* 411–424.

Krueger, R. F., Markon, K. M., Patrick, C. P., & Iacono, W. (2005). Externalizing psychopathology in adulthood: A dimensional-spectrum conceptualization and its implications for *DSM–V*. *Journal of Abnormal Psychology, 114,* 537–550.

Kuhn, T. S. (1962). *The structure of scientific revolutions.* Chicago: University of Chicago Press.

Levin, L., Bradley, R., & Westen, D. (2006). *Personality subgroups in adolescents with attention-deficit/hyperactivity disorder.* Manuscript submitted for publication.

Lingiardi, V., Shedler, J., & Gazzillo, F. (2006). Assessing personality change in psychotherapy with the SWAP-200: A case study. *Journal of Personality Assessment, 86,* 23–32.

Livesley, W. J., & Jackson, & D. N. (1992). Guidelines for developing, evaluating, and revising the classification of personality disorders. *Journal of Nervous and Mental Disease, 180,* 609–618.

Livesley, W. J., & Jackson, D. N. (in press). Manual for the Dimensional Assessment of Personalty Pathology–Basic Questionnaire. Port Huron, MI: Sigma Press.

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

Marin-Avellan, L., McGauley, G., Campbell, C., & Fonagy, P. (2005a, February). *Associations between violence and personality disorders using the SWAP-200.* Paper presented at the annual conference of the British and Irish Group for the Study of Personality Disorder, Glasgow, Scotland.

Marin-Avellan, L., McGauley, G., Campbell, C., & Fonagy, P. (2005b). Using the SWAP-200 in a personality-disordered forensic population: Is it valid, reliable and useful? *Journal of Criminal Behaviour and Mental Health, 15,* 28–45.

McGlashan, T. H., Grilo, C. M., Sanislow, C. A., Ralevski, E., Morey, L. C., Gunderson, J. G., et al. (2005). Two-year prevalence and stability of individual *DSM–IV* criteria for schizotypal, borderline, avoidant, and obsessive-compulsive personality disorders: Toward a hybrid model of Axis II disorders. *American Journal of Psychiatry, 162,* 883–889.

Meehl, P. E. (1954). *Clinical vs. statistical prediction.* Minneapolis: University of Minnesota Press.

Miller, J. D., & Pilkonis, P. A. (2006). Neuroticism and affective instability: The same or different? *American Journal of Psychiatry, 163,* 839–845.

Millon, T. (1991). Classification in psychopathology: Rationale, alternatives and standards. *Journal of Abnormal Psychology, 100,* 245–261.

Pilkonis, P. A., Heape, C. L., Ruddy, J., & Serrao, P. (1991). Validity in the diagnosis of personality disorders: The use of the LEAD standard. *Psychological Assessment, 31,* 46–54.

Robins, E., & Guze, S. (1970). The establishment of diagnostic validity in psychiatric illness: Its application to schizophrenia. *American Journal of Psychiatry, 126,* 983–987.

Russ, E., Bradley, R., Shedler, J., & Westen, D. (in press). Refining the narcissistic diagnosis: Defining criteria, subtypes, and endophenotypes. *American Journal of Psychiatry.*

Russ, E., Heim, A., & Westen, D. (2003). Parental bonding and personality pathology assessed by clinician report. *Journal of Personality Disorders, 17,* 522–536.

Sawyer, J. (1966). Measurement and prediction, clinical and statistical. *Psychological Bulletin, 66,* 178–200.

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.

Shedler, J., & Westen, D. (1998). Refining the measurement of Axis II: A Q-sort procedure for assessing personality pathology. *Assessment, 5,* 333–353.

Shedler, J., & Westen, D. (2004a). Dimensions of personality pathology: An alternative to the five factor model. *American Journal of Psychiatry, 161,* 1743–1754.

Shedler, J., & Westen, D. (2004b). Refining personality disorder diagnoses: Integrating science and practice. *American Journal of Psychiatry, 161,* 1350–1365.

Shedler, J., & Westen, D. (2007). The Shedler-Westen Assessment Procedure (SWAP): Making personality diagnosis clinically meaningful. *Journal of Personality Assessment, 89,* 41–55.

Waller, N. G., & Meehl, P. E. (1998). *Multivariate taxometric procedures: Distinguishing types from continua* (Vol. 9). Thousand Oaks, CA: Sage.

Watson, D., & Clark, L. A. (1992). Affects separable and inseparable: On the hierarchical arrangement of the negative affects. *Journal of Personality and Social Psychology, 62,* 489–505.

Westen, D. (2002). Clinical Diagnostic Interview. Unpublished manual, Emory University. Retrieved October 4, 2007, from www.psychsystems.net/lab

Westen, D. (1997). Divergences between clinical and research methods for assessing personality disorders: Implications for research and the evolution of Axis II. *American Journal of Psychiatry, 154,* 895–903.

Westen, D. (1998). The scientific legacy of Sigmund Freud: Toward a psychodynamically informed psychological science. *Psychological Bulletin, 124,* 333–371.

Westen, D., & Chang, C. (2000). Personality pathology in adolescence: A review. *Adolescent Psychiatry, 25,* 61–100.

Westen, D., Dutra, L., & Shedler, J. (2005). Assessing adolescent personality pathology: Quantifying clinical judgment. *British Journal of Psychiatry, 186,* 227–238.

Westen, D., Gabbard, G. O., & Blagov, P. (2006). Back to the future: Personality structure as a context for psychopathology. In R. F. Kreuger & J. L. Tackett (Eds.), *Personality and Psychopathology* (pp. 335–384). New York: Guilford Press.

Westen, D., & Harnden-Fischer, J. (2001). Personality profiles in eating disorders: Rethinking the distinction between Axis I and Axis II. *American Journal of Psychiatry, 165,* 547–562.

Westen, D., & Muderrisoglu, S. (2003). Reliability and validity of person-

ality disorder assessment using a systematic clinical interview: Evaluating an alternative to structured interviews. *Journal of Personality Disorders, 17,* 350–368.

Westen, D., & Muderrisoglu, S. (2006). Clinical assessment of pathological personality traits. *American Journal of Psychiatry, 163,* 1285–1287.

Westen, D., Nakash, O., Thomas, C., & Bradley, R. (2006). Clinical assessment of attachment patterns and personality disorder in adolescents and adults. *Journal of Consulting and Clinical Psychology, 74,* 1065–1085.

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

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

Westen, D., Shedler, J., & Bradley, R. (2006). A prototype approach to personality diagnosis. *American Journal of Psychiatry, 163,* 838–848.

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

Westen, D., Waller, N., Blagov, P., Shedler, J., & Bradley, R. (2007). *Measuring pathological personality traits by clinician-reports using the SWAP-II: Factor structure, validity, and retest reliability.* Unpublished manuscript.

Westen, D., & Weinberger, J. (2004). When clinical description becomes statistical prediction. *American Psychologist, 59,* 595–613.

Widiger, T. A., & Samuel, D. B. (2005). Evidence-based assessment of personality disorders. *Psychological Assessment, 17,* 278–287.

Wilson, T. D., Lindsey, S., & Schooler, T. Y. (2000). A model of dual attitudes. *Psychological Review, 107,* 101–126.

Wood, J. M., Nezworski, M. T., Lilienfeld, S. O., & Garb, H. N. (2003). *What's wrong with the Rorschach? Science confronts the controversial inkblot test.* San Francisco: Jossey-Bass.

Zanarini, M. C., Frankenburg, F. R., Hennen, J., & Silk, K. R. (2003). The longitudinal course of borderline psychopathology: 6-year prospective follow-up of the phenomenology of borderline personality disorder. *American Journal of Psychiatry, 160,* 274–283.

Received February 15, 2007
Revision received August 10, 2007
Accepted August 10, 2007 ■