

The Shedler-Westen Assessment Procedure (SWAP): Evaluating Psychometric Questions About Its Reliability, Validity, and Impact of Its Fixed Score Distribution

Pavel S. Blagov, Wu Bi, Jonathan Shedler and Drew Westen

Assessment published online 11 February 2012

DOI: 10.1177/1073191112436667

The online version of this article can be found at:

<http://asm.sagepub.com/content/early/2012/02/08/1073191112436667>

Published by:



<http://www.sagepublications.com>

Additional services and information for *Assessment* can be found at:

Email Alerts: <http://asm.sagepub.com/cgi/alerts>

Subscriptions: <http://asm.sagepub.com/subscriptions>

Reprints: <http://www.sagepub.com/journalsReprints.nav>

Permissions: <http://www.sagepub.com/journalsPermissions.nav>

>> [OnlineFirst Version of Record](#) - Feb 11, 2012

[What is This?](#)

The Shedler-Westen Assessment Procedure (SWAP): Evaluating Psychometric Questions About Its Reliability, Validity, and Impact of Its Fixed Score Distribution

Assessment
XX(X) 1–13
© The Author(s) 2012
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1073191112436667
http://asm.sagepub.com


Pavel S. Blagov¹, Wu Bi², Jonathan Shedler³, and Drew Westen⁴

Abstract

The Shedler-Westen Assessment Procedure (SWAP) is a personality assessment instrument designed for use by expert clinical assessors. Critics have raised questions about its psychometrics, most notably its validity across observers and situations, the impact of its fixed score distribution on research findings, and its test-retest reliability. We review empirical data addressing its validity, emphasizing the multitrait-multimethod approach to evaluating test validity. To evaluate the hypothesis that the fixed, asymmetric score distribution artifactually inflates correlations between SWAP profiles, we conducted Monte Carlo simulations and also presented empirical data from a large patient sample. We observed a mean correlation of zero between simulated SWAP profiles, indicating that the score distribution does not impact the correlation coefficients. Empirical correlations between SWAP profiles of actual patients were small and similar to those obtained using *Diagnostic and Statistical Manual of Mental Disorders*, fourth edition (DSM-IV) personality disorder scales that had no fixed score distributions, suggesting that the correlations were not a methodological artifact of the SWAP. We report new test-retest reliability data (median coefficient > .85) for the SWAP's trait and personality disorder dimensions. The SWAP appears to be reliable and valid. The data do not support its primary psychometric critiques.

Keywords

personality disorders, Shedler–Westen Assessment Procedure, SWAP-II, Q-sort, fixed distribution, test–retest reliability, validity

The Shedler–Westen Assessment Procedure (e.g., SWAP-II; Shedler & Westen, 2004a, 2004b; Westen & Shedler, 2007) is a measure of personality and personality pathology designed for use by clinically trained mental health professionals. Critics (Block, 2008; Widiger, 2002; Wood, Garb, Nezworski, & Koren, 2007) have raised a number of questions about the reliability and validity of the procedure and, more specifically, about particular psychometric properties of the scales and scores derived from SWAP data. In this article, we evaluate the three primary psychometric questions and criticisms of the SWAP identified in the literature, primarily pertaining to its performance as a tool for the dimensional assessment of personality pathology in research on the classification of personality pathology (psychiatric nosology).

Some critics (Wood et al., 2007) have questioned the validity of existing SWAP scales, raising legitimate concerns about the extent to which available research has examined their cross-method/cross-informant validity (Question 1).

We address this critique through a review of the existing literature, including very recent research.

Others have questioned the reliability and psychometric soundness of SWAP score profiles and scale scores more generally. Thus, another critique (Question 2) is the argument that the fixed score distribution of the instrument results in artifactually high correlations between SWAP-derived scale scores and personality pathology prototypes (Block, 2008) or artificially high estimates of reliability and validity (Wood et al., 2007). We address Question 2

¹Whitman College, Walla Walla, WA, USA

²UTC Power Corp, South Windsor, CT, USA

³University of Colorado, Aurora, CO, USA

⁴Emory University, Atlanta, GA, USA

Corresponding Author:

Pavel S. Blagov, PhD, Whitman College, 345 Boyer Avenue, Walla Walla, WA 99362-2067, USA

Email: blagovp@whitman.edu

conceptually as well as empirically in Studies 1 and 2 by examining hypothesized biases of the fixed-score distribution on SWAP profile intercorrelations in simulated data as well as in a large sample of real patients.

In addition, Wood et al. (2007) have drawn attention to the fact that published data on the test–retest reliability of SWAP scores have been lacking, a legitimate concern we address (Question 3) in Study 3.

The Shedler–Westen Assessment Procedure

The SWAP is a psychometric system designed to provide a comprehensive assessment of personality and personality pathology (Shedler & Westen, 2004a, 2004b, 2007; Westen & Shedler, 1999a, 1999b). Unlike most personality assessment instruments, the SWAP is neither a self-report questionnaire nor a rating form for lay informants. Rather, it is an instrument designed for use by trained mental health professionals in the context of either a thorough examination of a patient using a systematic clinical research interview (Westen & Muderrisoglu, 2003, 2006; <http://www.psychsystems.net/manuals>) or in a professional assessment or ongoing therapeutic engagement (e.g., longitudinal knowledge of the patient over the course of psychotherapy). In this sense, it resembles the Psychopathy Checklist–Revised (PCL-R; Hare, 2003; Hare & Neumann, 2006), which can be scored by a forensic professional from a research interview, from all available data (including chart review), or both.

The premise of the SWAP approach is that a clinically trained informant who has examined a patient over time or completed a systematic clinical research interview resembling the narrative-based interviewing process used by clinicians of all theoretical orientations in practice (Westen, 1997; Clinical Diagnostic Interview, www.psychsystems.net/manuals) can make reliable and valid observations and inferences about psychological processes that may not be accessible via self-report or readily observable by nonexperts (Westen & Weinberger, 2004). The SWAP instruments consist of 200 items, which the assessor sorts into eight categories, from *not descriptive* (0) to *most descriptive* (7) of the person. The procedure may be completed using paper cards with the items printed on them, an electronic spreadsheet designed to facilitate the rating and sorting process, or Internet interface programmed to serve the same purpose.

The SWAP instruments (the original SWAP-200 and revised SWAP-II for adults, and the SWAP-200-A and revised SWAP-II-A for adolescents) have been used to develop an empirically based classification of personality disorders (PDs; Shedler & Westen, 2007; Westen & Shedler, 1999a, 1999b; Westen, Shedler, Bradley, & DeFife, in press; Westen, Waller, Shedler, & Blagov, in press), to refine current diagnostic constructs by identifying richer diagnostic

criterion sets more faithful to the clinical syndromes observed in practice as well as in the lab (Blagov & Westen, 2008; Russ, Bradley, Shedler, & Westen, 2008; Shedler & Westen, 2004a; Zittel & Westen, 2005), to identify clinically important personality dimensions via factor analysis that are absent from other dimensional models of personality (Shedler & Westen, 2004b; Westen, Shedler, Bradley, & DeFife, in press), to link SWAP-assessed dimensions to etiological and outcome variables (including, e.g., genetic history, psychosocial history variables, and treatment response to both psychotherapy and pharmacological interventions; Westen & Shedler, 2007), to develop dimensional prototype models for personality diagnosis as an alternative to the categorical approach of *Diagnostic and Statistical Manual of Mental Disorders*, fourth edition (*DSM-IV*; Shedler & Westen, 2004a; Spitzer, First, Shedler, Westen, & Skodol, 2008; Westen, Shedler, & Bradley, 2006; Westen, Shedler, et al., in press), to explore subthreshold forms of personality pathology currently absent from the diagnostic manuals (Blagov, Bradley, & Westen, 2007), to assess subtypes of Axis I and Axis II disorders (e.g., DiLallo, Jones, & Westen, 2009; Russ et al., 2008), and to assess change in psychotherapy (Lingiardi, Shedler, & Gazillo, 2006).

The SWAP, like all instruments based on the Q-sort method, uses a fixed score distribution. In other words, assessors rank-order items for their degree of applicability to the patient at hand rather than rating them, and they must assign each rank or score a specified number of times (e.g., limiting the highest ranking scale points, in the case of the SWAP, a rank of 5, 6, or 7, defined as most descriptive of the patient, to a small number of items, while assigning lower scores to a higher number of items). Prior Q-sort instruments, such as the California Adult Q-Sort (Block, 1978) have treated items as bipolar dimensions (*very uncharacteristic* to *very characteristic*) and have used quasi-normal score distributions in which middle scores indicate neutrality on the dimension (e.g., Block, 1978; Shedler & Block, 1990), and hence are the most common ranks that can be assigned using the fixed distribution. In contrast, SWAP items assess unipolar constructs, and the fixed score distribution is therefore asymmetric, ranging from 0 (*not descriptive*) to 7 (*highly descriptive*). Many items receive scores of 0 and progressively fewer items receive higher scores. In other words, SWAP items are descriptive of a person to a greater or lesser degree, but they are not defined as *negatively* descriptive.

The rationale for this approach has been discussed at length elsewhere (Westen & Shedler, 2007), but three considerations were most important in establishing the distribution. First, empirically, virtually all psychopathology items from all scales are asymmetrically distributed in nature, with most patients showing little or no evidence of them and increasingly fewer individuals showing extreme scores. Thus, most people receive a score of 0 to 5 on the Beck

Depression Inventory (e.g., Beck, Brown, & Steer, 1996) or the PCL-R and a 0 on most of the items; progressively fewer receive scores above 25 on either instrument. Second, empirically, when we tested the first iteration of the SWAP by asking experienced clinicians to *rate* a patient in their care using a traditional Likert-type scale, this asymmetrical distribution was in fact what we obtained (Shedler & Westen, 1998). Finally, the meaning of many personality statements is unclear if they are defined as bipolar traits, such that a low score is ambiguous. For example, the opposite of having emotions that spiral out of control could be having emotions that are flat or having relatively normal affect experience and regulation. Similarly, the opposite of being chronically anxious could mean being unable to experience anxiety or having normal anxiety levels. The asymmetric distribution resolves that problem by having coders make fine-grained distinctions among items that are *descriptive* of the patient to varying degrees but not those that are not descriptive. This also saves coders considerable time, as they do not have to spend most of their time considering the exact placement of items of little relevance to a given patient, to maximize the utility of the instrument in everyday practice. The SWAP distribution requirements are thus as follows:

Score	0	1	2	3	4	5	6	7
Frequency	100	22	18	16	14	12	19	8

Critiques of the SWAP have focused on a number of concerns, of which three have been the most common and bear on the use of the SWAP method and SWAP-derived scores in nosological research on personality pathology: the ability of SWAP-based variables to predict cross-method/cross-informant criterion variables (Question 1), the extent to which the SWAP's fixed-score distribution may bias estimates of the extent to which the personality descriptions of individuals match PD prototypes and inflate reliability and validity estimates (Question 2), and the lack of evidence for temporal stability of SWAP scale scores (Question 3). Below, we address Question 1 through literature review, Question 2 conceptually and empirically (Studies 1 and 2), and Question 3 empirically (Study 3).

Question 1: Do SWAP Scales Show Cross-Method and Cross-Informant Validity?

The first critique (Wood et al., 2007) is based on the fact that much of the early research using the SWAP (e.g., Westen & Shedler, 1999a, 1999b) tended to rely on a single expert informant, the treating clinician, to provide both SWAP ratings and relevant criterion data (e.g., PD symptom ratings, adaptive functioning measures, and family history variables). This was indeed a legitimate concern.

The first studies that bear on this critique were small-sample studies finding very high correlations (in the range of $r = .70-.80$) between two independent clinical observers describing the same patient using the SWAP, one based on a systematic clinical interview and the other from the treating clinician, with each blind to the data and scoring of the other. These data resemble data on interrater reliability in that the scores being compared are based on observations generated by different evaluators describing the same patient; however, they also differ from interrater reliability estimates in that interrater reliability is traditionally based on comparing observations by different judges of the same or a similar sample of behavior, whereas the correlations obtained in these studies were based on different observers evaluating completely different samples of behavior by the same target individual (e.g., data from a single cross-sectional interview and data from longitudinal observation of the patient in treatment over months). These correlations can thus be interpreted as yielding evidence for validity to the extent that one informant, usually the one who completed a formal clinical-research assessment of the patient blind to all available data, is held as the predictor rater, whereas the other informant, usually the treating clinician, whose observations are based on detailed observation of and interaction with the patient over time, is held as the criterion rater.

For example, in a sample of 24 outpatients, personality syndromes calculated from independent interviewer ratings had a median correlation of .80 with the same SWAP dimensions calculated from the treating clinicians' assessments of the patients. Equally relevant to assessing their validity was the impressive discriminant validity of the SWAP dimensions when tested against *DSM-IV* constructs, personality prototypes derived empirically, and personality traits (Westen & Muderrisoglu, 2003, 2006). These findings have just been replicated in a large-sample study, in which 145 patients were evaluated independently by a systematic clinical research interview (based on a single cross-sectional encounter with the patient, as is the norm in PD research) and by the treating clinician (based on all available longitudinal data from ongoing treatment), with each coder blind to the data from the other. Convergent validity correlations averaged .50 and discriminant validity correlations hovered around 0.0 (Westen, Shedler, et al., in press).

In a sample of 30 inpatients at a maximum-security forensic hospital conducted by an independent research team, 12 SWAP-based PD scales derived from researchers' ratings of attachment interviews and chart records correlated highly with interpersonal circumplex ratings of patients' personalities provided by their nurses as well as whether the patients had committed violent offenses (Marin-Avellan, McGauley, Campbell, & Fonagy, 2005). This work has recently been extended to a larger sample of 90 participants (Marin-Avellan, 2010).

In a study of 47 outpatients (Bradley, Hilsenroth, Guarnaccia, & Westen, 2007), SWAP dimensions of Borderline, Antisocial, and Obsessive–Compulsive PDs based on ratings by treating clinicians showed a strong pattern of concurrent and discriminant validity with relevant dimensions of the Personality Assessment Inventory (PAI; Morey, 1991), a self-report measure of personality pathology completed by the patients. Correlations between SWAP dimensions and conceptually overlapping PAI scales ranged from .31 to .46, showing cross-method and cross-informant validity levels similar to those found in the literature using other established self-report instruments (e.g., Clifton, Turkheimer, & Oltmanns, 2005). Evidence of discriminant validity was strong, with no correlations even trending toward significance between the two near-neighbor disorders (BPD and APD) from the SWAP and the PAI and significant negative correlations in the range of $r = -.35$ –.40 between both disorders as assessed by the SWAP and Obsessive–Compulsive PD as assessed by the PAI.

Yet another team (Smith, Hilsenroth, & Bornstein, 2009) reported cross-method/cross-informant evidence for the validity of the SWAP-200 dependency scales evaluated against *DSM-IV* PD diagnoses by clinicians and self-report using an interpersonal circumplex measure in a sample of 85 patients. Whereas the studies by Westen and Muderrisoglu (2003, 2006) and Westen and colleagues (Westen, Shedler, et al., in press; Westen, Waller, et al., in press) offer increasingly strong evidence for cross-informant validity, the studies by Bradley et al. (2007), Marin-Avellan et al. (2005), and Smith et al. (2009) offer evidence from three different, independent, research teams, both for cross-informant and for cross-method validity of specific SWAP scales.

In a sample of 91 highly psychopathic (PCL-R > 39) male inmates at medium- and high-security units, Blagov et al. (2011) used the SWAP in conjunction with videotaped clinical interviews to derive two psychopathy subtypes empirically. The primary psychopathy subtype was distinguished by malevolent narcissistic features and high levels of superficial, seductive, and deceitful charm, whereas the secondary psychopathy subtype was emotionally dysregulated, impulsive, and hostile, akin to Borderline PD (American Psychiatric Association; 2000). The extent to which participants' SWAP profiles matched the empirical primary and secondary psychopathy prototypes evidenced good convergent validity with PCL-R factor scores obtained by an independent set of raters, with self-report measures of personality and psychopathology completed by the participants, and with observer-report measures of psychopathy-related interpersonal behavior and impulsivity. The participants' degrees of match to the two SWAP-based psychopathy prototypes evidenced excellent discriminant validity with regard to the external validation measures.

For example, primary psychopathy correlated with PCL-R Factor 1 (from an independent team of assessors; $r = .45$), a self-report measure of extraversion ($r = .41$), a self-report measure of positive emotionality ($r = .60$), a self-report scale of active temperament ($r = .42$), a self-report measure of neuroticism ($r = -.37$), a self-report measure of fearful temperament ($r = -.50$), and a self-report scale of internalized anger ($r = -.35$). Secondary psychopathy had no significant correlations with PCL-R Factor 1, extraversion, active temperament, and neuroticism. It correlated significantly with PCL-R Factor 2 ($r = .27$), records of antisocial and violent behavior during childhood and childhood abuse ($r = .40$ –.44), self-report anger expression ($r = .45$) and internalized anger ($r = .42$), positive affectivity (inversely, $r = -.60$), socialization ($r = -.55$), observer-report aggression ($r = .57$) and inattention/hyperactivity ($r = .36$), and self-report indices of negative emotionality ($r = .42$ –.52). Thus, two SWAP-II scales capturing psychopathy subtypes evidenced strong cross-method and cross-informant validity.

Published data thus suggest that the first critique of the SWAP, namely regarding the cross-informant/cross-method validity of its scales, requires revision. Multiple research groups have now shown in multiple samples strong evidence for validity using multitrait–multimethod designs. Similar findings have emerged whether the SWAP studies used dimensional *DSM-IV* PD scales, empirically derived SWAP trait scales, or empirically derived personality configurations or PD prototypes, with highly similar estimates of validity regardless of which kind of scales were used and how the scales were computed (e.g., Q-correlations vs. traditional unit-weighted scales, a point to which we return in addressing Question 2). At this point, few studies are being conducted using a single clinical informant, in part because of the widespread currency of this criticism and in part because the goals of the original studies that relied on large samples of patients as described by individual clinicians participating in practice networks were taxonomic, required prohibitively large samples for cross-informant research, and have largely been fulfilled (Westen, Shedler, et al., in press; Westen, Waller, et al., in press). The validity evidence for cross-informant/cross-method data, has, however, generally been of similar magnitude to previously published validity data using a single informant, the treating clinician, suggesting that the use of a single informant did not in fact impose substantial bias.

Of note, researchers have generally not held self-report personality instruments and structured interviews that rely heavily or exclusively on patient self-reports to this same standard, often accepting at face value correlations between self-reported personality pathology and self-reports of other variables interpreted as demonstrating validity. In general, personality researchers would do well to return to evaluating validity using a multitrait/multimethod matrix approach that includes cross-correlations across informants.

Question 2: Does the Asymmetrical Fixed Distribution Generate Biases and Artificially Inflated Estimates of Validity and Reliability?

The SWAP score distribution has been challenged by proponents of other models of personality assessment, notably the five-factor model (Widiger, 2002; Widiger & Samuel, 2005; Wood et al., 2007). Two related critiques have received the most attention. The first is that the asymmetrical distribution of the SWAP may artificially inflate correlations between individual SWAP profiles and criterion prototypes used to generate scale scores (Block, 2008). These correlations are also known as Q-scores (i.e., correlations between a given patient's 200-item profile and an empirically derived, aggregated 200-item profile, e.g., of BPD or APD). Block points out that the fixed distribution forces a large number of items (100 items or 50%) into the "not descriptive" category (items that receive a rank of 0), and he argues that the sheer number of items ranked 0 would inflate Q-scores even though they are not "highly descriptive" items (i.e., items receiving ranks of 4, 5, 6, or 7).

Westen and Shedler (2007) have addressed this and similar concerns elsewhere; however, because of the persistence of this critique, it is worth briefly addressing it here conceptually before addressing it empirically. Conceptually, the assertion that the score distribution induces spuriously high correlations is mathematically untenable. Because of the fixed distribution, the items in a SWAP profile always have a mean of 1.69 and a standard deviation of 2.18. This is also true of empirically derived SWAP-based prototypes of personality pathology that have previously been used to compute Q-scores. Thus, in computing a correlation between a SWAP profile and a prototype (or between two SWAP profiles), items with scores from 0 to 3 deviate minimally from the mean item score and therefore can have only minimal impact on the magnitude of the correlation coefficient. In effect, the top 44 items (those receiving scores from 4 to 7, which are 2 or 3 to 5 standard deviations higher than the mean) "drive" the magnitude of the correlation coefficient (because no item can be substantially lower than the mean, rendering items with low scores largely irrelevant to the magnitude of the correlation). This is consistent with the unipolar nature of the items, where items receiving a high ranking are especially *descriptive* of the patient. The large number of items with scores of 0 mathematically has minimal impact on the magnitude of a Pearson correlation coefficient, given its computation, which is derived from deviation of each score from the mean. Thus, should features of some kind of personality pathology that is not central to understanding the patient be present among even items ranked 1 to 3 (but not among items ranked 4-7), the deviation from the mean of 1.69 will have minimal impact

on the magnitude of Q-correlations. For items receiving a 0, the impact is trivial.

Were this critique valid, there would also be a substantial discontinuity between the results of SWAP research relying on Q-correlations (a scaling method derived from Block, 1978) to measure personality constellations, such as PDs and research using factor-analytically derived SWAP trait scales, which are traditional unit-weighted item scales. That has not in fact been the case with the SWAP-200 (Shedler & Westen, 2004a, 2004b), the SWAP-200-A for adolescents (Westen, Dutra, & Shedler, 2005), or the SWAP-II (Westen, Waller, et al., in press). Unit-weighted trait scales have produced similar patterns of convergent and discriminant validity as PD scales derived from Q-correlations.

In any case, whether this criticism about the application of Q-scores to a given patient has merit conceptually or mathematically, it is no longer relevant, given recent changes in the scaling of SWAP scores for both *DSM-IV* and empirically derived personality syndromes (Westen, Shedler, et al., in press), whether for purposes of clinical assessment or research. In developing SWAP-II scales to measure *DSM-IV* dimensional diagnoses, we tested multiple methods for generating scales and discovered that Q-correlations were actually less valid than traditional scales ranging from 16 to 24 items, depressing correlations between SWAP PD scales and measures of Axis II pathology while inflating estimates of comorbidity—precisely the opposite of the criticism suggested by Wood et al. (2007), Block (2008), and others. The differences were not enormous, but they were substantial enough that we have turned to more conventional scaling techniques that take the average of items relevant to the construct, rather than correlating a patient's 200-item profile with a 200-item aggregate profile except where the latter proves empirically more predictive or useful. Thus, this first version of the criticism, even if it were true, is now essentially moot.

Critics have, however, expressed concerns about other possible undesirable effects of the asymmetrical fixed distribution of the SWAP (Block, 2008; Wood et al., 2007), particularly as it is or has been used for research purposes, notably on estimates of reliability and validity. These critiques are similar to the ones addressed above, resting on the large number of items receiving a ranking of 0 and otherwise low scores in the asymmetrical distribution. Block argued that "all SWAP-II intercorrelations are appreciably heightened because of the inevitable and appreciable overlap of heavily weighted and inapplicable descriptors" so that "the chance correlation between SWAP Q-sorts begins at a surprising, unwittingly high level" (Block, 2008, p. 114). Were this true, of course, it would inflate estimates of discriminant validity as much as convergent validity (because all correlations would increase) and hence would not have any net effect on estimates of validity or reliability

to a psychometrician appropriately attending to correlations both on and off the diagonal.

In the previous section, we addressed why, conceptually and mathematically, low-ranking items are, in fact, not weighted heavily and have minimal impact on correlations between SWAP profiles. Furthermore, it has been known in the Q-sort literature for some time that “the exact distribution form has little effect on the kinds of analyses which are made of the data [because] correlation coefficients, and the factors obtained from them, are largely insensitive to changes in distribution shapes” (Nunnally, 1978, p. 616). Nevertheless, the critique merits an empirical evaluation, both because it has been raised frequently by critics and because it is possible for the fixed distribution to have some effect on correlation magnitudes if a subset of the items were consistently to receive low ratings (especially 0; Block, 2008) or receive very different mean ranks and standard deviations (Nunnally & Bernstein, 1994). As described elsewhere, however, items with minimal variance or extremely low base rates have been deleted from the SWAP during its construction as part of a multi-year, iterative revision process (Westen & Shedler, 2007).

Here, we present two studies that address this question. To evaluate the critique empirically, we considered the following: (a) What is the average chance correlation among SWAP profiles? We addressed this in Study 1. In Study 2, we address two additional questions: (b) Do empirical data support the claim of high “baseline” correlations between pairs of SWAP profiles in an actual patient sample, and are these higher than the baseline correlations between, for example, profiles of patients assessed using *DSM-IV* criteria rated on a similar but unconstrained Likert-type rating scale? (c) What is the typical SWAP item score and its dispersion? In other words, is it true that some SWAP items consistently receive scores of 0 across individuals and others consistently receive high scores, thereby “building in” a spurious correlation between score profiles?

Study 1 is a Monte Carlo simulation in which we examine the magnitude of correlations between simulated SWAP profiles created by generating random scores conforming to the fixed distribution requirement of the SWAP. If the asymmetric fixed distribution artifactually inflates correlations between SWAP profiles, we should observe positive correlations between the simulated SWAP profiles, and the Monte Carlo procedure will permit us to estimate the magnitude of the inflation. In Study 2, we examine empirically observed correlations among SWAP profiles of actual patients in a large national sample of patients assessed with the SWAP-II (Westen & Shedler, 2007), comparing them with correlations observed using unconstrained 6-point ratings (no fixed distribution) of Axis II diagnostic criteria. We also report central tendency and dispersion estimates for the individual SWAP-II items to explore whether certain items have score distributions that might bias estimates of reliability and validity.

Study 1: Monte Carlo Simulation of Randomly Paired SWAP Profiles

We wrote a computer program to generate 10,000 random SWAP profiles, each of which conformed to the fixed distribution requirements described above (e.g., with only 8 items allowed to receive a score of 7 and 100 items required to receive a score of 0). We then selected random pairs of SWAP profiles from this sample and calculated the correlations between the profile pairs, resulting in 5,000 correlation coefficients. We calculated the mean, median, standard deviation, standard error of the mean, and skewness of the 5,000 correlations. To generate a sampling distribution of correlation coefficient means, we repeated this entire procedure 200 times.

The mean correlation coefficients were normally distributed (as expected under the central limit theorem), centered on a mean of $M = .001$ ($SD = .001$, $SEM < .001$, $Md = .001$). The Shapiro–Wilk test of normality ($SW = .992$, $df = 200$, $p = .395$) was not significant. The distribution of the skewness statistic ($M = .077$, $SD = .033$, $SEM = .002$, $Md = .079$) was similarly approximately normal ($SW = .993$, $df = 200$, $p = .463$), indicating that the distributions of correlations had minimal skew. The standard error of skewness (a constant equal to .035 for this sample size) suggested a slight positive skew to the distribution of correlations, although the normality plots for the mean and skewness were consistent with approximate normality with small deviations at the extremes.

The data thus show that in 200 samples of 5,000 SWAP correlations, the mean correlation was 0 to 2 decimal places, with a negligible positive skew. The sample size we chose (5,000) was 5 to 10 times the size of the largest existing studies on personality pathology and roughly 50 times the size of the average study. In research with real participants, even in the largest-N studies, the slight deviation from normality in the sampling distribution would thus have no impact on research findings, and the mean and median correlations among simulated patient pairs of 0.0 do not support critics’ speculations.

Study 2: SWAP Profile Intercorrelations in a Clinical Sample

The aim of the next study was to evaluate empirically the critique that the fixed distribution may inflate profile intercorrelations by examining (a) intercorrelations of SWAP-II profiles of real patients compared with intercorrelations among profiles using an alternative (*DSM*-based) measure; and (b) central tendencies and variability estimates of item scores across patients.

Method

We used data collected as part of a larger study whose methods and overall sample characteristics have been detailed

elsewhere (e.g., Shedler & Westen, 2007; Westen & Shedler, 2007). In brief, we invited psychologists and psychiatrists from a random national sample (30% response rate; $N = 1,201$) to serve as expert informants and offered them a \$200 consultation fee to describe “an adult patient you are currently treating or evaluating who has enduring pattern of thoughts, feeling, motivation or behavior—that is, personality problems—that cause distress or dysfunction.” We explicitly informed potential clinician-participants that patients did not need to have a PD diagnosis. Patients were required to be at least 18 years old, not in a psychotic episode, and well-known to the clinician (at least 6 hours but less than 2 years of clinical contact, to minimize effects of personality change with treatment). To limit selection biases, we directed observers to select the last patient they saw during the previous week who met study criteria. Of a subsample of clinicians recontacted with questions regarding the patient they selected, 96% reported following the procedure as specified. Each observer contributed data on only one patient to minimize rater-dependent variance.

The expert informants completed the SWAP-II along with additional measures of psychopathology, adaptive functioning, personal and family history, and other clinically relevant information. Among these materials and of relevance to the present study was an Axis II Criteria Rating Form (a randomly ordered list of *DSM-IV* diagnostic criteria for all 10 PDs). Clinician-informants rated each diagnostic criterion on a scale from 1 to 6, with values greater than 3 indicating that the criterion was definitely met. The scale is similar in range to the 0 to 7 scale used in the SWAP, except that it does not have a fixed distribution and hence provides a useful psychometric comparison. By applying *DSM-IV* decision rules to the criterion data (coded as criterion present or absent), we were also able to assign categorical Axis II diagnoses to patients.

Results and Discussion

Intercorrelations Among SWAP-II Profiles in Real Patients

To evaluate SWAP profile intercorrelations among the patients, in a first analysis we included those individuals who met *DSM-IV* diagnostic criteria for at least one Axis II PD based on Axis II diagnostic criteria. To reduce true covariance due to the presence of overlapping personality features, we excluded pairs of patients with Axis II diagnoses falling within the same Axis II diagnostic cluster (e.g., cluster A, which subsumes the comorbid diagnoses of paranoid, schizoid, and schizotypal PDs). We predicted that we would observe symmetrically distributed correlations with values ranging from negative to positive, with the average correlation positive and small in magnitude (because most individuals sampled from a clinical population with PD

diagnoses share certain aspects of personality pathology features, such as negative affectivity, and because of non-negligible correlations between dimensional diagnoses across clusters in all prior research using structured interviews, self-reports, the SWAP, and other instruments).

For comparison purposes, we also generated a distribution of correlations for the same patient subsample using an alternative *DSM*-based measure, the Axis II Criterion Rating Form, that did not have a fixed score distribution. We predicted that the distribution of the SWAP-II data would resemble that of the *DSM-IV* criterion ratings, except that the SWAP-II data would have a somewhat higher mean because the SWAP-II includes symptoms related to general neurotic features (e.g., tends to be anxious, tends to be depressed) that have been excluded from the *DSM-IV* to minimize comorbidity but are shared nearly universally among patients with PDs. Finally, we repeated these two analyses, except that this second time we used the entire sample of 1,201 patients (i.e., including those without a PD) and calculated all possible SWAP-II and *DSM-IV* Axis II profile intercorrelations (without restricting the possible pairings to nonoverlapping PD clusters). We expected that, in these latter analyses, both the SWAP-II and the *DSM-IV* Axis II Criterion Rating data would produce mean intercorrelations between random pairs of real patients that are small to moderate in magnitude and similar to one other.

The first analysis resulted in 89,128 unique correlations between pairs of SWAP-II profiles, which were distributed symmetrically around a mean of .07386 ($SEM = .00055$, $Md = .07614$) with $SD = .16454$. The skewness statistic (.00186) was smaller than its standard error (.00821), suggesting the absence of significant skewness. The correlations between Axis II criterion ratings were similarly distributed symmetrically around a mean of .01510 ($SEM = .00065$, $Md = .01576$) with $SD = .19478$. The skewness statistic (.02315) was greater than twice the standard error (.00821), suggesting a small but statistically significant positive skew.

We interpret these results as follows. When we sampled pairs of patients who did not share PDs from the same Axis II cluster, the measure that was the basis of this selection (the Axis II Criterion Rating Form) produced a roughly normal but positively skewed distribution of correlations that was centered slightly above zero. The SWAP-II produced slightly larger correlations that on average were nevertheless small. As noted above, one of the meaningful ways in which the SWAP-II and the Axis II Criterion Ratings differ as measures of personality pathology is that the former measures general symptoms of distress and personality dysfunction related to negative affectivity or neuroticism that were not included among the *DSM-IV* criteria to reduce comorbidity; if a tendency to experience negative emotions such as depression were included among the PD criteria, they would have to be included among the criteria for multiple disorders because empirically they are related to several disorders (e.g., Shedler & Westen, 2004a, 2004b).

In fact, in the present sample, the following SWAP-II items had the highest mean ranks across patients: Tends to feel unhappy, depressed, or despondent ($M = 4.39$); Tends to feel anxious ($M = 4.12$); Tends to fear s/he will be rejected or abandoned ($M = 3.95$); and Tends to feel he or she is inadequate, inferior, or a failure ($M = 3.82$). Thus, in light of the evidence from Study 1, and in the absence of substantial skewness, the shift in the mean of the SWAP-II distribution of correlations likely reflects the fact that patients sampled from a clinical population who are in treatment for personality pathology share a proclivity toward anxiety, depression, rejection sensitivity, and low self-esteem. In the context of interrater reliability estimation, this covariance probably reflects the accurate perceptions of raters and their agreement rather than an artifactual bias having to do with the distributional properties of the measure. Indeed, it may reflect a likely strength of the SWAP-II, which was intended to measure the broad spectrum of personality pathology, relative to the roughly 80 items comprising the diagnostic criteria for the PDs in *DSM-IV*. In any case, if the marginal difference of .05 between data using the fixed distribution and the data using a nonfixed distribution is meaningful and replicable for whatever reasons, it affects convergent and discriminant validity alike and hence has no bearing on validity or on reliability as long as discriminant reliability estimates are obtained along with convergent reliability estimates, something we recommend in PD diagnosis given substantial comorbidities among disorders.

In examining the SWAP-II correlation data when the patients in each pair had not been preselected to eliminate patients from the same *DSM* diagnostic cluster, 720,600 correlations were distributed unimodally and symmetrically around a mean of $M = .152$ ($SD = .171$, $SEM < .001$, $Md = .152$), with a skewness statistic of $-.012$ that was smaller in absolute value than its standard error of .003. The 713,415 *DSM-IV* Axis II Criterion Rating patient profile correlations were distributed around a virtually identical mean of $M = .143$ ($SD = .202$, $SEM < .001$, $Md = .148$) with a skewness of $-.133$ that was in absolute value substantially larger than its standard error of .003. Although the two correlation means were significantly different from each other ($p < .01$) in this extraordinarily large sample, the 95% confidence interval = .009-.010 indicated that the real mean difference was negligible.

These data suggest that the SWAP-II (which has a fixed, asymmetric distribution of items) and the Axis II criterion rating form (which uses 6-point scale ratings and imposes no restrictions on the score distribution) produced similar distributions of random intercorrelations among patients with personality pathology not constrained to those with an Axis II PD. On average, nearly identical, small, positive correlations (.14 and .15) were observed with both instruments, suggesting that people who seek treatment for personality problems have on average 2% shared variance in

their personality profiles when assessed with either personality pathology measure.

Central Tendency and Variability of SWAP-II Ratings

We computed central tendency (means, medians, and modes) and variability (standard deviations) estimates for the SWAP-II items in the entire participant sample to examine the extent to which some items may tend to appear in extreme locales of the distribution and thus might unduly influence profile intercorrelations.

The overwhelming majority of SWAP-II items (196) had a modal score of 0, with the exception of the following four items: Item 35 (“Tends to feel anxious,” Mode = 7), Item 92 (“Is articulate, can express self well in words,” Mode = 7), Item 98 (“Tends to fear she or he will be rejected or abandoned,” Mode = 6), and Item 189 (“Tends to feel unhappy, depressed, or despondent,” Mode = 7). The scores given to each of the 200 items ranged from 0 to 7, showing that each item received scores from the full possible range. The average mean score was 1.69 ($SD = 0.85$). The median scores per item ranged from 0 to 5, with an average median of .98 ($SD = 1.19$), suggesting that the tendency was for the majority of the items to receive low scores. Item standard deviations ranged from 1.10 to 2.50 ($M = 1.97$, $SD = 0.33$). These data suggest that, with the exception of a few items that tended to be highly descriptive of most patients, the majority of the items “hover” at the bottom of the score distribution and receive high rankings only when clinicians find those rankings appropriate. This does not support the claim (Block, 2008) that some select items might have undue influence because their range is restricted to the bottom of the distribution. Furthermore, with the exception of the same few items noted above, the variability of the means of the items’ ranks is constrained within the lower end of the fixed-score distribution and the standard deviations of the items’ ranks do not vary substantially. This reduces the concern that unequal item means and standard deviations may inflate profile correlations (Nunnally & Bernstein, 1994), although the effect of the four items that had much higher means needs to be evaluated in future research. The results can be seen in Table 1.

Question 3: Do SWAP Scales Have Reasonable Test–Retest Reliability?

As noted by Wood et al. (2007), evidence for test–retest reliability (temporal stability) of the SWAP has been lacking. Recently, an independent group (Cogan & Porcerelli, 2011) reported that, on reevaluation of 77 patients with the SWAP-200 after a 6-month retest interval, reliabilities averaged $r = .81$ for the instrument’s PD scales and $r = .68$ for its empirically derived personality diagnostic syndromes.

Table 1. Summary Statistics of Central Tendency and Variability Estimates of 200 SWAP-II Items in a Sample of 2,101 Adult Patients

Statistic	Summary statistics for 200 items						
	N	Range	Minimum	Maximum	M	SEM	SD
Item M	200	4.04	0.3	4.34	1.69	0.06	0.85
Item Md	200	5	0	5	0.98	0.08	1.19
Item mode	200	7	0	7	0.14	0.07	0.95
Item SD	200	1.40	1.10	2.51	1.97	0.02	0.33
Item range	200	0	7	7	7	0	0
Item minimum	200	0	0	0	0	0	0
Item maximum	200	0	7	7	7	0	0

This study fills an important void in the literature; however, the question of temporal stability is a legitimate concern for any instrument that cannot be adjudicated with a single study. Thus, in Study 3, we present the first data on the test–retest reliability of factor-analytically derived trait dimensions of the SWAP-II in a clinical sample, complementing the data provided by Cogan and Porcerelli, as well as data on PD scores, as reported by Cogan and Porcerelli.

Study 3: Test–Retest Reliability of the SWAP-II

The aim of the study was to address Question 3 by examining the test–retest reliability of the SWAP factors and PD scales, thus addressing a gap identified in the literature (Wood et al., 2007). As part of the project referenced in Study 2 above, we recontacted 139 participating informants who had provided SWAP-II data using the Internet approximately 2 to 4 months earlier and asked them to assess their patient again. At retest, 40 received the SWAP-II paper card-sort to complete, 40 received a username and password to complete the instrument on the Internet, and 59 received the prior version of the instrument, the SWAP-200 (to provide preliminary data useful for imputation from one to the other). The response rate was 68% ($N = 27$, paper, $N = 29$, Internet, and $N = 38$, SWAP-200; total $N = 94$). Retest data arrived 4 to 6 months following the initial assessment. The retest subsamples did not differ at $p = .05$ in the discipline, gender, and years of experience of the observers, or in the gender, age, GAF (Global Assessment of Functioning), or time in treatment of the patients.

We correlated scores for the 14 factor-analytically derived SWAP-II trait scales (Westen, Waller, Shedler, & Blagov, in press) as well as 10 PD scales (based on the *DSM-IV* criteria as represented in SWAP-II items) at Time 1 and Time 2, which allowed us to assess not only test–retest reliability along the diagonal but also discriminant reliability estimates off the diagonal, something that has largely been

lacking in the PD literature (Clark, 1992). To examine cross-method stability (paper vs. Internet), we calculated reliability matrices separately for observers who completed both assessments using the Internet and those who completed one paper and one web version, although with the small sample sizes we consider these analyses very preliminary. We used Fisher's Z to test for any differences between web-to-paper and web-to-web correlations. We chose a study-wise alpha level of $p < .01$, given the large number of tests for which we made no predictions. To maximize N for retest analyses, we used simple algorithms to impute Time 2 SWAP-II profiles for patients whose clinicians had used the SWAP-II at Time 1 and the SWAP-200 at Time 2. Because the SWAP-II items correlated highly with corresponding SWAP-200 items, only those items that were completely new or substantially rewritten required imputation (16 items).

Results and Discussion

The clinician retest sample ($N = 94$) consisted of 81% psychologists and 19% psychiatrists. Most self-identified theoretically as integrative or eclectic (52%), cognitive–behavioral (18%), or psychodynamic (18%). Patients had a mean age of 42.4 years ($SD = 13.0$) and were roughly 65% male. The patients were receiving treatment primarily in private practice (62%), outpatient clinics (20%), or inpatient settings (10%). The ethnic distribution was 74% Caucasian, 11% African American, 10% Hispanic, and 5% other ethnicities.

Table 2 contains the results for the factor-analytically derived scales. In Westen, Waller, et al. (in press), we derived 16 factors, which showed generally good internal consistency with a mean α of .73, although two (Boundary Disturbance and Sexual Conflict) were on the low side at .44 and .55, most likely because of a limited number of highly loading items (4 and 6, respectively). We retained 14 of the 16 factors, eliminating Boundary Disturbance because of its low internal consistency and an eating disorder factor that we interpreted primarily as an Axis I scale, but we report data on all 16 factors here). Retest correlations for the 16 SWAP-II factors (along the diagonal) ranged from .64 to .96 (all significant at $p < .001$), with a median retest $r = .85$. The only factor with a reliability coefficient below .70 was Factor 13 (Boundary Disturbance), which had also shown relatively weak internal consistency. Correlations off the diagonal (between each factor at Time 1 and all other factors at Time 2) averaged near 0 and were distributed with a slight negative skew, $M = -.04$, $SD = .26$. Medians of the absolute values of the discriminant retest correlations for each factor were substantially lower than the test–retest correlations and ranged from $r = .08$ to .32 (for Obsessionality). Interestingly, cross-method (web-to-paper) retest reliability coefficients were slightly higher than same-method (web-to-web) coefficients, though not significantly, $M_d = .87$ and .78, respectively. Thus, the two

Table 2. Test–Retest Correlations for 16 SWAP-II Scales ($N = 94$)

	f1r	f2r	f3r	f4r	f5r	f6r	f7r	f8r	f9r	f10r	f11r	f12r	f13r	f14r	f15r	f16r	Md_{abs}^a
f1	.96*	-.38*	-.56*	.09	-.10	.26	.47*	-.24	-.14	-.31*	-.53*	.55*	.26	.07	.51*	-.24	.26
f2	-.43*	.84*	.20	-.36*	-.01	-.30*	.01	-.07	-.30*	-.22	-.10	-.36*	-.01	.05	-.31*	-.08	.20
f3	-.51*	.10	.84*	.13	.40*	-.39*	-.27*	.06	.17	.13	.35*	-.46*	-.22	-.29*	-.28*	-.01	.27
f4	.06	-.32*	.11	.81*	.42*	-.09	-.16	.08	-.09	-.01	.23	.03	-.17	-.39*	.10	-.10	.10
f5	.01	.04	.22	.20	.77*	-.48*	.06	-.18	-.03	-.06	-.06	-.14	-.13	-.38*	.06	-.20	.13
f6	.24	-.37*	-.39*	.02	-.39*	.87*	-.02	.17	.06	.09	-.08	.26	.16	.18	.29*	-.02	.17
f7	.53*	.07	-.35*	-.20	-.01	.01	.90*	-.32*	-.22	-.46*	-.73*	.11	.46*	.19	.43*	-.22	.22
f8	-.28*	-.16	.20	.01	-.18	.19	-.30*	.85*	.09	.32*	.39*	-.09	-.19	-.12	-.25	.15	.19
f9	.01	-.31*	.08	.01	-.01	.01	-.14	-.07	.70*	.08	.23	.08	-.04	.04	-.19	.24	.08
f10	-.27*	-.24	.07	.05	-.04	.07	-.43*	.32*	.24	.85*	.45*	-.12	-.24	-.18	-.21	.21	.21
f11	-.54*	-.16	.42*	.26	.02	-.14	-.68*	.28*	.34*	.47*	.86*	-.19	-.33*	-.18	-.49*	.28*	.28
f12	.57*	-.36*	-.46*	.07	-.25	.30*	.05	-.09	-.06	-.17	-.18	.86*	.10	.11	.12	-.15	.15
f13	.31*	-.08	-.32*	-.06	-.22	.37*	.40*	-.01	-.14	-.25	-.39*	.14	.64*	.27*	.26	-.19	.25
f14	.14	.10	-.36*	-.39*	-.47*	.13	.29*	-.17	.06	-.17	-.30*	.12	.26	.83*	-.17	.12	.17
f15	.54*	-.27*	-.35*	.17	.01	.34*	.34*	-.14	-.31*	-.22	-.40*	.19	.15	-.15	.92*	-.18	.22
f16	-.20	-.04	-.01	-.14	-.14	-.01	-.16	.14	.19	.16	.20	-.13	-.19	.13	-.16	.87*	.14
Md_{abs}^a	.28	.16	.32	.13	.14	.19	.27	.14	.14	.17	.30	.14	.19	.18	.25	.18	.85

Note. SWAP-II = Shedler–Westen Assessment Procedure–II; f1 = Psychopathy; f2 = Psychological health; f3 = Obsessionality; f4 = Schizotypy; f5 = Emotional avoidance; f6 = Emotionally dysregulated; f7 = Narcissism; f8 = Anxious somatization; f9 = Sexual conflict; f10 = Depression; f11 = Social anxiety/avoidance; f12 = Unstable commitments; f13 = Boundary confusion; f14 = Histrionic sexualization; f15 = Hostility; f16 = Eating disturbance; r = Retest.

a. Median absolute value of the discriminant retest correlations.

* $p < .01$ (two-tailed).

methods appear to be equivalent. Nor did the imputed scores produce divergent patterns from the web-to-web and web-to-paper; hence their inclusion in these analyses.

Table 3 reports the results for the SWAP *DSM-IV* PD scales, using the current scaling procedures, which do not rely on Q-correlations. Prior research using the SWAP has tended to report results using these scales or their equivalents in the earlier version of the instrument (the SWAP-200). As Table 3 shows, the retest correlations (along the diagonal) were very high and ranged from .86 to .96 ($Md = .90$). The discriminant (off the diagonal) test–retest correlations were somewhat higher and had a larger standard deviation ($M = .08$, $SD = .49$) than the equivalent correlations for the factors in Table 1. The absolute values of the discriminant retest correlations for the PD scales ranged from .01 to .80, with a median of .37. Thus, the SWAP-II factors had similar test–retest reliabilities but lower artifactual “comorbidities” when compared with the SWAP-II PD scales, which makes sense given that the factors were empirically derived using factor analysis and hence do not suffer from the problem of comorbidity built into all measures of *DSM-IV* constructs.

In summary, we examined the test–retest reliability of 16 SWAP-II factors and PD scales over a 4- to 6-month period in a subsample of 94 patients treated in a wide range of clinical settings. Within the limits of sample size, the results indicate substantial retest reliability. Whereas the 16 SWAP-II factors yielded high correlations with themselves over 4 to 6 months ($Md_r = .85$), the median correlation

across factors was modest, providing evidence for discriminant retest reliability, something that is too infrequently reported in studies of personality pathology measures where constructs are often overlapping. The SWAP-II PD scales were similarly reliable but more “comorbid,” underscoring the general consensus in the literature on PDs that the *DSM-IV* PDs show unrealistically high comorbidity estimates.

Three limitations are worth noting vis-à-vis the retest data. The first reflects the unintended oversampling of men that took place as a result of the timing of the retest study in relation to the overall study. Second, temporal stability beyond 4 to 6 months is unknown. We chose this follow-up interval to minimize personality change reflecting treatment as a potential confound. A third limitation is the reliance on a single informant, which means that part of the stability in the psychopathology factors may reflect rater effects, although this is an equal limitation of all retest reliability data on self-report scales.

Summary and Conclusions

We evaluated the three most prevalent psychometric critiques in the literature concerning the SWAP. Regarding the validity critique (Question 1), the literature review found substantial evidence for cross-informant/cross-method validity of SWAP scales measuring dimensional *DSM-IV* diagnoses, empirically derived PDs, and traits, as well as preliminary data on primary and secondary

Table 3. Test–Retest Correlations for 10 DSM-IV PD Scales of the SWAP-II (N = 94)

	1	2	3	4	5	6	7	8	9	10	Md _{abs} ^a
1 Paranoid PD	.92*	-.04	.25*	.59*	.45*	.35*	.54*	-.45*	-.43*	-.22	.25
2 Schizoid PD	-.12	.87*	.71*	-.11	-.37*	-.52*	-.32*	.53*	.01	.32*	.11
3 Schizotypal PD	.17	.75*	.88*	.09	-.16	-.28*	-.05	.24	-.15	.25*	.09
4 Antisocial PD	.60*	-.12	.08	.96*	.42*	.50*	.65*	-.57*	-.50*	-.45*	.08
5 Borderline PD	.39*	-.26*	-.09	.40*	.89*	.68*	.21	-.29*	.03	-.58*	.03
6 Histrionic PD	.33*	-.51*	-.29*	.53*	.72*	.90*	.52*	-.59*	-.16	-.62*	.16
7 Narcissistic PD	.63*	-.35*	-.08	.70*	.22	.49*	.94*	-.80*	-.66*	-.18*	.08
8 Avoidant PD	-.55*	.54*	.28*	-.61*	-.36*	-.58*	-.78*	.92*	.62*	.29	.36
9 Dependent PD	-.46*	.05	-.04	-.44*	.02	-.10	-.54*	.56*	.88*	-.13*	.10
10 Obsessive–compulsive PD	-.25*	.27*	.16	-.40*	-.58*	-.55*	-.16	.25*	-.11	.86*	.16
Md _{abs} ^a	.17	.04	.08	.09	.02	.10	.05	.29	.15	.18	.90

Note. DSM-IV = *Diagnostic and Statistical Manual of Mental Disorders* fourth edition; PD, personality disorder; SWAP-II = Shedler–Westen Assessment Procedure–II.

a. Median absolute value of the discriminant retest correlations.

* $p < .01$ (two-tailed).

psychopathy. Research underway with a sample of more than 200 patients is examining the relation between SWAP DSM-IV diagnoses, empirically derived PDs, and factor-analytically derived traits with a range of criterion variables, from data from multiple independent interviewers and self-reports, to genotyping, to adaptive functioning, to longitudinal follow-up data at 18 months.

Regarding the distributional critiques (Question 2), the fixed, asymmetric distribution of the SWAP has been a particular source of controversy. Although we have moved away from Q-correlations as a method of scaling, particularly for personality assessment in practice, neither conceptually nor empirically does this critique appear to be well grounded. Study 1 demonstrated that the average intercorrelation between randomly generated SWAP profiles, conforming to the asymmetric fixed distribution requirement, is zero to two decimal places. In Study 2, using real data from a national sample of $N = 1,201$ patients with personality pathology, the correlations between random pairs of patient profiles were essentially equal, on average, whether using the SWAP-II with its fixed distribution of scores (average $r = .15$) or a set of DSM-IV-based Axis II Criterion Ratings that did not have a fixed distribution (average $r = .14$). These average values are not 0, but they do not likely reflect the fixed score distribution of the SWAP, given that they are essentially identical for 6-point unconstrained DSM-IV criterion ratings. They are more likely to reflect naturally occurring similarities in patients who seek therapy and who have personality pathology, as evidenced by the fact that four SWAP-II items (reflecting proneness to anxiety, depression, rejection sensitivity, and high verbal skills) were the only four items whose median score across 1,201 patients was not 0 (and was, in fact, high). Recently, in a sample of psychotherapy patients (Cogan & Porcerelli,

2011), the SWAP-200 scale scores of *mismatched* patients (i.e., pairs consisting of unrelated patients at Time 1 and Time 2, instead of the same patient at test and retest) correlated at $r = .02$ (for PD scales) and $r = -.05$ (for Q-factor analysis-derived scales), also suggesting that reliability coefficients are not artificially inflated. The fact that, in Study 2, the mean and median ratings for 196 out of 200 items were close to 0 (and that the mean and variability estimates of the *standard deviations* were small) argues against Block's (2008) concern that a sizeable proportion of items consistently receives low ratings whereas another sizable proportion consistently receives high ratings.

Regarding Question 3, that temporal stability data had been lacking, we reported the second test–retest study of the SWAP and the first evidence for test–retest reliability of SWAP traits derived by factor analysis (Study 3), demonstrating an average $r = .85$ in an adult clinical sample over a 4- to 6-month period. The mean reliability was .90 for SWAP scales corresponding to DSM-IV personality disorders. Research is needed on the retest reliability of these factors over a longer period, preferably in a population that is not undergoing treatment for personality pathology.

In sum, we evaluated three of the primary questions critics have raised about the validity and psychometrics of scales based on the SWAP. Although critics can and will no doubt generate other concerns about the SWAP, as they should about any psychometric instrument, the data reviewed and presented here did not support these critiques. Other concerns, however, need to be addressed, namely with respect to questions confronting all personality instruments, particularly those that focus on psychopathology, such as the best approaches to scaling psychopathological constructs that have substantial but differing and relatively low base rates in the population. Like others who have been down this

same path, we have attempted to identify an optimal if imperfect way of transcending the limits of traditional but readily interpretable metrics, such as *T*-scores, which can provide a distorted impression of profiles of disorders with different base rates (e.g., Borderline vs. Schizotypal PD). We are now using normalized *T* scores (Hsu, 1984), which have the advantage of equating percentiles associated with *T* scores across traits, disorders, or other personality constellations.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by National Institute of Mental Health grant MH60892-01 (to D.W.).

References

- American Psychiatric Association. (2000). *The diagnostic and statistical manual of mental disorders* (4th ed., Text Rev.). Washington, DC: Author.
- Beck, A. T., Brown, G., & Steer, R. A. (1996). *Beck Depression Inventory II manual*. San Antonio, TX: Psychological Corporation.
- Blagov, P. S., 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.
- Blagov, P. S., Lilienfeld, S. O., Patrick, C. J., Powers, A. D., Phifer, J. E., Venables, N., . . . Cooper, G. (2011). Personality constellations in incarcerated psychopathic men. *Personality Disorders: Theory, Research, and Treatment, 2*, 293-315.
- Blagov, P. S., & Westen, D. (2008). Questioning the coherence of Histrionic Personality Disorder: Personality subtypes in adults and adolescents. *Journal of Nervous and Mental Disease, 169*, 785-797.
- Block, J. (1978). *The Q-sort method in personality assessment and psychiatric research*. Palo Alto, CA: Consulting Psychologists Press.
- Block, J. (2008). *The Q-sort in character appraisal: Encoding subjective impressions of persons quantitatively*. Washington, DC: American Psychological Association.
- Bradley, R., Hilsenroth, M., Guarnaccia, C., & Westen, D. (2007). Relationship between clinician assessment and self-assessment of personality disorders using the SWAP-200 and PAI. *Psychological Assessment, 19*, 225-229.
- Clark, L. (1992). Resolving taxonomic issues in personality disorders: The value of larger scale analyses of symptom data. *Journal of Personality Disorders, 6*, 360-376.
- Clifton, A., Turkheimer, E., & Oltmanns, T. F. (2005). Self- and peer perspectives on pathological personality traits and interpersonal problems. *Psychological Assessment, 17*, 123-131.
- Cogan, R., & Porcerelli, J. H. (2011). Test-retest reliability and discriminant validity of the SWAP-200 in a psychoanalytic treatment sample. *Psychology & Psychotherapy: Theory, Research & Practice*. doi: 10.1111/j.2044-8341.2011.02020.x
- DiLallo, J. J., Jones, M., & Westen, D. (2009). Personality subtypes in disruptive adolescent males. *Journal of Nervous and Mental Disease, 197*, 15-23.
- Hare, R. D. (2003). *The Hare Psychopathy Checklist-Revised* (2nd ed.). Toronto, Ontario, Canada: Multi-Health Systems.
- Hare, R. D., & Neumann, C. S. (2006). The PCL-R assessment of psychopathy: Development, structural properties, and new directions. In C. J. Patrick (Ed.), *Handbook of psychopathy* (pp. 58-88). New York, NY: Guilford Press.
- Hsu, L. (1984). MMPI T-scores: Linear versus normalized. *Journal of Consulting and Clinical Psychology, 52*, 821-823.
- Lingiardi, V., Shedler, J., & Gazillo, F. (2006). Assessing personality change in psychotherapy with the SWAP-200: A case study. *Journal of Personality Assessment, 86*, 23-32.
- Marin-Avellan, L. (2010). *Validation of the SWAP-200 with forensic patients: Do structuring clinical judgements of PD aid violence risk assessments?* (Unpublished doctoral dissertation). University College, London, England.
- Marin-Avellan, L., McGauley, G., Campbell, C., & Fonagy, P. (2005). Using the SWAP-200 in a personality-disordered forensic population: Is it valid, reliable, and useful? *Criminal Behaviour and Mental Health, 15*, 28-45.
- Morey, L. C. (1991). *The Personality Assessment Inventory: Professional manual*. Odessa, FL: Psychological Assessment Resources.
- Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.
- Nunnally, J. C., & Bernstein, I. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.
- Russ, E., Bradley, R., Shedler, J., & Westen, D. (2008). Refining the construct of narcissistic personality disorder: Diagnostic criteria and subtypes. *American Journal of Psychiatry, 165*, 1473-1481.
- Shedler, J., & Block, J. (1990). Adolescent drug use and psychological health: A longitudinal inquiry. *American Psychologist, 45*(5), 612-630.
- Shedler, J., & Westen, D. (1998). Refining the measurement of Axis II: A Q-sort procedure for assessing personality pathology. *Assessment, 5*, 335-355.
- Shedler, J., & Westen, D. (2004a). Refining personality disorder diagnoses: Integrating science and practice. *American Journal of Psychiatry, 161*, 1350-1365.
- Shedler, J., & Westen, D. (2004b). Dimensions of personality pathology: An alternative to the Five Factor Model. *American Journal of Psychiatry, 161*, 1743-1754.
- Shedler, J., & Westen, D. (2007). The Shedler-Westen Assessment Procedure (SWAP): Making personality diagnosis clinically meaningful. *Journal of Personality Assessment, 89*, 41-55.

- Smith, S. W., Hilsenroth, M. J., & Bornstein, R. F. (2009). Convergent validity of the SWAP-200 dependency scales. *Journal of Nervous and Mental Disease, 197*, 613-618.
- Spitzer, R. L., First, M. B., Shedler, J., Westen, D., & Skodol, A. E. (2008). Clinical utility of 5 dimensional systems for personality diagnosis: A "consumer preference" study. *Journal of Nervous and Mental Disease, 196*, 356-374.
- 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., Dutra, L., & Shedler, J. (2005). Assessing adolescent personality pathology: Quantifying clinical judgment. *British Journal of Psychiatry, 186*, 227-238.
- Westen, D., & Muderrisoglu, S. (2003). Assessing personality disorders using a systematic clinical interview: Evaluation of an alternative to structured interviews. *Journal of Personality Disorders, 17*, 351-369.
- Westen, D., & Muderrisoglu, S. (2006). Clinical assessment of pathological personality traits. *American Journal of Psychiatry, 163*, 1285-1287.
- 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. (2007). Personality diagnosis with the Shedler-Westen Assessment Procedure (SWAP): Integrating clinical and statistical measurement and prediction. *Journal of Abnormal Psychology, 116*, 810-822.
- Westen D., Shedler, J., & Bradley R. (2006). A prototype approach to personality disorder diagnosis. *American Journal of Psychiatry, 163*, 846-856.
- Westen, D., Shedler, J., Bradley, B., & DeFife, J.A. (in press). An empirically derived taxonomy for personality diagnosis: Bridging science and practice in conceptualizing personality. *American Journal of Psychiatry*.
- Westen, D., Waller, N., Shedler, J., & Blagov, P. (in press). Dimensions of personality and personality pathology: Factor structure of the Shedler–Westen Assessment Procedure–II (SWAP-II). *Journal of Personality Disorders*.
- Westen, D., & Weinberger, J. (2004). When clinical description becomes statistical prediction. *American Psychologist, 59*, 595-613.
- Widiger, T. A. (2002). Personality disorders. In M. M. Antony & D. H. Barlow (Eds.), *Handbook of assessment and treatment planning for psychological disorders* (pp. 453-480). New York, NY: Guilford Press.
- Widiger, T., & Samuel, D. B. (2005). Evidence-based assessment of personality disorders. *Psychological Assessment, 17*, 278-287.
- Wood, J. M., Garb, H. N., Nezworski, M. T., & Koren, D. (2007). The Shedler–Westen Assessment Procedure-200 as a basis for modifying DSM personality disorder categories. *Journal of Abnormal Psychology, 116*, 823-836.
- Zittel, C., & Westen, D. (2005). Borderline personality disorder as seen in clinical practice: Implications for DSM-V. *American Journal of Psychiatry, 162*, 867-875.