Journal of Personality Disorders, 29, 2015, 231
© 2015 The Guilford Press

DIAGNOSTIC AGREEMENT BETWEEN CLINICIANS AND CLIENTS: THE CONVERGENT AND DISCRIMINANT VALIDITY OF THE SWAP-200 AND MCMI-III PERSONALITY DISORDER SCALES

Emanuela S. Gritti, PhD, Douglas B. Samuel, PhD, and Margherita Lang, PsyD

A particularly controversial aspect in the field of personality assessment is the use of self-report measures, versus clinicians' evaluations, for diagnosing personality disorder (PD). No studies have systematically documented the agreement between these sources for the entire array of *DSM-5* PDs using comprehensive measures and experienced clinicians' judgments. The present work fills this gap by indexing the agreement between patients' self-descriptions and clinicians' judgments, relying on standardized and thorough PD instruments. The Shedler-Westen Assessment Procedure–200 (SWAP-200; Westen & Shedler, 1999a, 1999b) and the Millon Clinical Multiaxial Inventory-III (Millon, Davis, & Millon, 1997) were both completed in a clinical series of 56 adult outpatients. Analyses highlighted moderate correlations between the two measures for the 10 *DSM-5* PDs (*Mdn* = .35). Agreement was highest for psychological features that are more easily observable by the clinicians. Furthermore, results revealed problematic discriminant validity between the two instruments.

The diagnosis and assessment of personality disorders (PDs) has traditionally been a thorny issue that has created complications. Within practice settings, PD diagnoses are primarily assigned by treating therapists based on their clinical interactions and unstructured interviews with the patient (Perry, 1992; Westen, 1997). Nonetheless, the past three decades have seen a proliferation

From Department of Psychology, Milan Bicocca University, Milan, Italy (E. S. G., M. L.); Department of Psychological Sciences, Purdue University, West Lafayette, Indiana (D. B. S.); and Associazione per la Ricerca in Psicologia Clinica - A.R.P., Milan, Italy (M. L.).

We deeply thank the patients, clinicians, and staff of the Associazione per la Ricerca in Psicologia Clinica (A.R.P., Milan, Italy) where this study was conducted for the time and knowledge they shared with us. We also acknowledge all the colleagues with whom we had the opportunity to productively discuss over the years about the main topics discussed in this work.

Margherita Lang served as a co-editor for the Italian version of the MCMI-III and as such receives royalties for its use. No funding was received for this research.

Address correspondence to Emanuela S. Gritti, Department of Psychology, Milan Bicocca University, Piazza dell'Ateneo Nuovo 1, Milan 20126, Italy; E-mail: emanuela.gritti@unimib.it

of instruments designed to collect information directly from patients, and these self-report questionnaires or semistructured interviews have become the predominant method of PD assessment for empirical research (Lenzenweger, Loranger, Korfine, & Neff, 1997). Given these methodological discrepancies in the way PDs are diagnosed in practice and research, it is crucial to understand the degree to which these two sources of information overlap.

For example, treatment outcome research almost exclusively relies on diagnoses derived from interviews or self-report questionnaires. In this way, the translation of these empirical findings into evidence-based practice is *predicated* on the notion that the individuals receiving the diagnoses in practice and research settings are similar. Thus, it is critical to understand how PD ratings from both sources converge (and diverge) in order to determine how evidence-based treatments might apply to individuals in clinical settings. Although challenging to collect, diagnostic ratings from clinicians in the course of their own practice represent crucial data for moving the field forward. Such ratings have high external validity, particularly compared to ratings by research clinicians, who typically assign diagnoses based on a single clinical interview rather than clinical contact. This does not suggest that diagnoses assigned by researchers after diagnostic interviews are inaccurate or invalid—in fact, research points to the contrary (Samuel et al., 2013)—but merely that they do not provide information about real-world practices. Collecting such research is complicated by the time it requires from already busy mental health professionals and thus relies either on generosity on the part of the clinicians or significant compensation (e.g., $200; Westen, Shedler, Bradley, & DeFife, 2012).

Existing research has demonstrated that the correspondence between clinicians' ratings and self-report questionnaires completed by the patients is weak. For example, in one study, the median Kappa agreement between self-report questionnaires and clinician diagnoses for individual PDs was .08 (Hyler, Rieder, Williams, & Spitzer, 1989). Agreement between patient-report and unstructured clinician ratings was only moderately better when PD diagnoses have been considered dimensionally. The median cross-method correlations for individual PD diagnoses have ranged from a low of .05 (Chick, Sheaffer, Goggin, & Sison, 1993) to a high of.36 (Klein et al., 1993).

Westen and Weinberger (2004) have argued, however, that those studies have not provided a precise estimate of agreement because the source of information (i.e., the clinician) was conflated with the method of collection. Indeed, the clinicians' PD diagnostic ratings in the studies just noted have been confined to existing chart diagnoses or collected via brief rating forms. In contrast, self-report questionnaires typically contain hundreds of items that systematically and comprehensively assess aspects of each PD. Thus, the calculation of clinicians' agreement with self-report questionnaires is potentially suppressed by the problematic psychometric properties of the instruments used to collect them.

Shedler and Westen developed a more systematic tool aimed at allowing clinicians to quantify their clinical judgments. The Shedler-Westen Assessment Procedure (SWAP-200; Westen & Shedler, 1999a, 1999b) consists of 200 statements that were designed to capture important aspects of personality pathology, including *DSM-IV* PDs. Following the Q-sort method (Block, 1961), the clinician must identify 100 items that are *not descriptive* of the individual and then place the remaining 100 items into seven progressively smaller piles using a fixed distribution.

Researchers have criticized the SWAP-200 on the basis of its fixed, skewed distribution; the unrepresentative normative sample from which it was developed; and possible idiosyncrasies of its T-score approach to diagnosis (Wood, Garb, Nezworski, & Koren, 2007). Others have questioned the intrinsic subjectivity implied in clinicians' ratings of patients' personalities through the SWAP (Michels, 2012). Empirical findings, however, have demonstrated that SWAP-200 scores relate to concurrently assessed criterion variables (Bradley, Jenei, & Westen, 2005), and SWAP-200 ratings have obtained stronger agreement among clinicians than more routine methods (e.g., Westen & Muderrisoglu, 2003).

Only a few studies have explored whether SWAP descriptions improve the convergence of clinicians' diagnoses with self-report measures completed by patients (Bradley, Hilsenroth, Guarnaccia, & Westen, 2007; Davidson, Obonsawin, Seils, & Patience, 2003). Davidson and colleagues (2003) modified the SWAP-200 to make it suitable for use as a self-report inventory (e.g., rephrasing items that were clinical observations and eliminating items that could not be completed by self-report, such as "shows evidence of unconscious homosexual wishes or interests"). Twenty-three patients and their treating clinicians completed this modified SWAP-200 instrument, and the intraclass correlations (ICCs) for the PD Q-factors ranged from −.10 (depressive) to .67 (dysphoric) with a median value of .28, leading Davidson and colleagues to conclude that "the low ICC ($r < .07$) indicates poor agreement for the prototype scales" between clinician and patient ratings (p. 215). Although the Davidson et al. (2003) study was useful in indexing the agreement between the SWAP-200 and a self-report measure, it was limited by the self-report version of the SWAP that was specifically modified for the study. Although the use of an identical inventory for both sources is an intriguing experimental manipulation, there are clear advantages to using a commonly used self-report inventory.

Bradley and colleagues (2007) partially addressed this limitation by comparing the convergence between SWAP-200 PD scales, completed by the treating clinician, and the Personality Assessment Inventory (PAI; Morey, 1991). Eighteen graduate students described a total of 47 patients using the SWAP-200, and the patients completed the PAI. The SWAP-200 borderline PD scale correlated moderately with the PAI, including .31 with borderline features, .40 with affective stability, and −.07 with self-harm. Similarly, the SWAP-200 antisocial PD scale correlated .35 with PAI antisocial features, .21 with stimulation seeking, .45 with aggressive attitudes, and .46 with drug problems. Bradley and colleagues (2007) concluded that these "moderate to small correlations" were "consistent with the prior literature on self-informant cross-correlations" (p. 228).

Although useful for addressing a concern about the study by Davidson et al. (2003), Bradley and colleagues' (2007) findings also have limitations and leave important questions unanswered. Specifically, the PAI includes scales directly relevant to only two of the *DSM-5* PDs, and even these scales represent features of borderline and antisocial PDs rather than the exact diagnostic constructs. Utilizing a self-report inventory designed to explicitly assess all the *DSM-5* PDs would allow the calculation of formal cross-method convergence statistics. In addition, the study by Bradley, Hilsenroth, and colleagues (2007) utilized a clinical sample that was relatively large for clinician ratings (*N* = 47), but the SWAP-200 was completed by relatively inexperienced student therapists within a training clinic. We hypothesize that more experienced therapists might provide more valid SWAP-200 ratings. In sum, no prior study has examined the agreement between SWAP-200 ratings, assigned by clinicians in the course of their practice, and a self-report measure of the full complement of the *DSM-5* PDs. This knowledge is crucial for understanding the validity of PD diagnoses assigned in clinical practice.

The current study investigates the convergence between the PD scales from the SWAP-200, completed by therapists based on their clinical interactions with a patient, and the self-reported Millon Clinical Multiaxial Inventory–III (MCMI-III; Millon, Davis, & Millon, 1997). The MCMI-III is well suited for such a comparison because it enjoys wide use as a self-report measure of the *DSM-5* PDs. On the basis of the limited available literature on clinicians' ratings and their overlap with other methods, we hypothesize that the PD scales from the SWAP-200 and MCMI-III will correlate between .30 to .40. This estimate is comparable to the median correlation (*r* = .36) reported across 10 studies that compared PD ratings completed by self-report and peer informants (Klonsky, Oltmanns, & Turkheimer, 2002).

METHOD

PARTICIPANTS AND PROCEDURES

The present study involved outpatients of a private mental health clinic located in Milan, Italy. All patients undergo a diagnostic process (composed of psychological testing via a standard assessment battery and history taking) soon after intake in order to evaluate psychological functioning and plan treatment. After this evaluation, some patients start treatment whereas others spend a few sessions with the clinician to identify and discuss some key features of their psychological functioning that emerged from the integrative assessment.

The present data concern a clinical series of 59 patients who underwent this assessment process. Of the 59 patients included in the series, only those with valid MCMI-III records were selected (*n* = 56). MCMI-III records that included more than 12 invalid (i.e., responses both with a "yes" and a "no") or omitted responses or that obtained a Validity Index (i.e., Scale V: items 65, 110, and 157) score > 1 were considered invalid. To maximize the external validity, there were no eligibility criteria, and inclusion depended solely on whether the referring clinician completed the SWAP-200. Of the clinicians contacted to participate in the study, almost all consented and completed the SWAP-200. All patients provided written, informed consent as part of routine clinic procedures to indicate that their de-identified data could be used for research purposes. The group of patients included 34 females and 22 males, with a mean age of 35.2 years (*SD* = 11.5). They were largely White (98%), but included one Asian. This sample included a variety of educational levels, with four having completed 8 years of schooling. The average number of years of education was 15.5 (*SD* = 3.3), with a mode of 18 years.

The treating clinicians (*n* = 16) were predominantly female (69%) and White, with a mean age of 57.2 years (*SD* = 10.9). Clinicians were quite experienced, with an average of 29 years of practice (*SD* = 12.6). Most clinicians had a degree in psychology (44%) or analogous title within the Italian education system (31.3%), and 25% of them had a degree in medicine. All clinicians had a qualification equivalent of a doctoral degree in the American education system (e.g., 4–5 years of education and practicum training), and three of them were also members of the International Psychoanalytical Association. With respect to theoretical orientation, half indicated a psychodynamic orientation, 31% psychoanalytic, 13% cognitive-behavioral, and 6% systemic. These therapists were also experts on personality assessment as reflected by the fact that a majority (81%) of the clinicians had advanced training in identifying, through a multimethod assessment and extended history taking, the focus of the patient's psychological functioning conceived of as a potentially powerful therapeutic factor.

The procedure largely relied on the standard routine of the clinic, providing a naturalistic, externally valid test of our hypotheses. Clinicians had contact with their patients as part of their regular practice, including about 2–3 intake sessions before referring them to the assessment process, which was undertaken with a licensed psychologist who specialized in psychological assessment. Importantly for the present study, the MCMI-III was administered to clients during this assessment process (i.e., separately from the treating clinician who completed the SWAP-200). We further ensured that the MCMI-III records were kept in separate folders and not attached to the assessment reports delivered to the clinicians, maintaining independence between the self-report and clinician-report ratings. Later in the treatment process, the primary clinician provided ratings on the SWAP-200.

As a result of this natural setting, there was not a rigidly predetermined interval between the moment when the patient took the MCMI-III and the rating of the SWAP-200 by the clinician. In most cases (*n* = 29), the SWAP-200 was completed between 2 and 12 months later, but some were rated within the first 2 months (*n* = 11), and some were rated after more than a year (*n* = 19). The overall median was 6.0 months after the completion of the MCMI-III. Although this duration has considerable benefit because it indicates that clinicians would have been quite well acquainted with the patient, it does introduce a temporal component that potentially complicates a direct cross-method comparison. Using the three subdivided groups described above (≤ 2 months; 2–12 months; ≥ 12 months), we recalculated the cross-method correlational matrix for each group. Thus, we could determine whether the convergence between SWAP-200 and MCMI-III differed based on time interval.

Finally, another factor that potentially complicates our analyses is that clinicians, who rated multiple clients, might skew agreement. The number of clients rated by each of the 16 clinicians ranged from one to nine, with a median of three clients per therapist. We controlled for a potential "clinician effect" on the SWAP-200 by separately computing the convergent correlations between the MCMI-III and the SWAP after removing one clinician at a time from the total sample and analyzing the resulting matrices. In this way, we can determine if individual clinicians notably impact the agreement between the two instruments.

MATERIALS

Shedler-Westen Assessment Procedure–200 (SWAP-200). The SWAP-200 (Westen & Shedler, 1999a, 1999b) is an assessment rated by an observer with knowledge of the individual, normally the clinician, to describe personality. Its items encompass both specific behaviors (e.g., Item 40: "Tends to engage in unlawful or criminal behavior") and more inferential processes (e.g., Item 76: "Manages to elicit in others feelings similar to those he or she is experiencing; e.g., when angry, acts in such a way as to provoke anger in others; when anxious, acts in such a way as to induce anxiety in others"). The standard version of the SWAP-200 was adopted in this study; clinicians were therefore asked to evaluate patients using the software version of the SWAP-200 that allows one to electronically "sort" the 200 cards into the identified piles for ease of collection and scoring the data. The SWAP-200 ratings were then scored for the *DSM-5* Personality Disorders (PD T-Scores) because they are most directly comparable to the scales from the MCMI-III.

Because the data collection took place in Italy and all the clinicians spoke Italian, they completed the Italian version of the SWAP (Westen, Shedler, & Lingiardi, 2003). The Italian translation of the items was done by Vittorio Lingiardi and Francesco Gazzillo in collaboration with a work group of the Società Psicoanalitica Italiana, Centro Milanese di Psicoanalisi, and the SWAP's original authors. The internal consistencies of the SWAP-200 PD scales were above .90 in the development sample of the instrument (Westen & Shedler, 1999a). Nonetheless, due to the Q-sort nature of these scales, the alpha values are calculated using the patients as items, so they are not directly comparable to values from traditional instruments. The Italian version has been used widely in process and outcome research (e.g., Lingiardi, Shedler, & Gazzillo, 2006) as well as on group studies with a variety of clinical populations and measures (Gazzillo et al., 2013).

Millon Clinical Multiaxial Inventory-III (MCMI-III). The MCMI-III (Millon, Davis, & Millon, 1997) is a self-report personality inventory that consists of 175 items in a True/False format. The MCMI-III is based on Millon's conceptualization of personality, but most importantly for the present study, it provides scores for the 10 *DSM-5* PDs. The MCMI-III raw scores are transformed into weighted base rate (BR) scores that are used for interpretation purposes. The MCMI-III is widely used as a measure of personality with a broad literature of support (e.g., Barbot, Hunter, Grigorenko, & Luthar, 2013). The Italian adaptation of the MCMI-III, produced through a translation and back-translation process and approved for use by Pearson Assessment Inc., has been used in prior studies (e.g., Zennaro et al., 2013). The internal consistency of the scales in the Italian validation sample was greater than .80 for all 20 scales.

RESULTS

DESCRIPTIVE STATISTICS

The descriptive statistics for the 10 *DSM-5* PDs from the SWAP-200 (*PD T-Scores)* and MCMI-III (base rates) are provided in Table 1. Notably, the skewness and kurtosis values for all the variables fall within an acceptable range, indicating the relative normality of the distribution.

DIMENSIONAL AGREEMENT BETWEEN CLINICIAN AND SELF-REPORTED PD

Given that the temporal duration between the clients' self-reports and the therapists' ratings varied considerably across the cases, a first step in our analyses was to determine if this had an effect on the agreement between the two sources. Correlation matrices were computed separately for the three levels of the "time between MCMI-III and SWAP-200" nominal variable. The median convergent correlation was .33 when the SWAP-200 was completed less than 2 months after the MCMI-III, .23 when the interval fell between 2

TABLE 1. Descriptive Statistics for Target Variables (*n* = 56)

	Min	Max	Mean	Standard deviation	Skewness	Kurtosis	Number of PD diagnoses in the sample
SWAP-200 VARIABLES							
Paranoid	29.1	68.63	42.9	8.4	0.68	0.20	1
Schizoid	33.4	68.2	46.6	8.0	0.52	0.02	4
Schizotypal	31.63	70.97	46.4	8.5	0.45	0.01	3
Antisocial	38.02	62.14	46.8	6.8	0.55	–0.98	1
Borderline	31.7	64.9	46.6	7.4	0.11	–0.13	3
Histrionic	35.24	63.19	49.7	5.8	–0.13	0.36	2
Narcissistic	35.01	66.16	47.0	7.8	0.75	–0.32	5
Avoidant	29.8	62.6	46.5	7.8	0.01	–0.59	2
Dependent	32.05	62.81	49.4	7.8	–0.10	–0.82	4
OCPD	34.28	063.53	47.9	6.8	0.02	–0.52	2
MCMI-III VARIABLES							
Paranoid	0	81	37.0	26.5	–0.22	–1.48	0
Schizoid	0	101	49.7	20.7	–0.42	–0.17	1
Schizotypal	0	99	43.0	26.2	–0.50	–0.85	2
Antisocial	8	115	52.7	24.1	0.04	0.14	4
Borderline	0	111	46.7	27.9	–0.05	–0.62	4
Histrionic	0	108	59.2	23.0	–0.26	0.04	5
Narcissistic	6	98	57.8	20.0	–0.19	–0.35	5
Avoidant	0	97	49.7	27.3	0.07	–1.20	8
Dependent	0	107	54.8	27.8	–0.37	–0.71	7
OCPD	0	115	58.3	24.2	–0.02	0.15	6

and 12 months, and .31 when the interval was longer than 12 months. Given the relative invariance of agreement across the three levels of the time variable, we focus our interpretation on the overall correlation matrix. However, full matrices for each time interval are available by contacting the first author.

We calculated an overall multitrait-multimethod correlation matrix to index the degree of dimensional agreement between scores on the SWAP-200 and MCMI-III PD. Given the limited base rates for a variety of PDs, this dimensional approach provided a more accurate reflection of the agreement in this outpatient setting. Table 2 shows the full matrix of the correlations between SWAP-200 scores and MCMI-III scores for all 10 PDs. However, to facilitate a more streamlined presentation of these findings, Table 3 summarizes the results of this correlation matrix by focusing on the values for convergent (i.e., those along the diagonal of the lower left quadrant) and aggregated discriminant validity.

Convergent Validity. The magnitude of correlation between the self- and clinician-reported scores varied considerably across the PDs, ranging from a low of –.10 (paranoid) to .45 (avoidant), with a median of .35. In addition to the negative value for paranoid PD, the agreement for schizotypal was also minimal ($r = .05$), but all remaining values were .25 or higher, with six of them considered medium effect sizes according to Cohen (1992).

Clinician Effect. To investigate the possibility that a single clinician systematically biased our results, we also recalculated the correlation matrix after removing each clinician, sequentially. Analogous to the practice of calculating corrected item-total correlations, this strategy allowed us to isolate the impact of each clinician. Table 4 summarizes the minimum and maximum convergent correlations for each PD when removing one clinician at a time. A comparison between these values and the overall convergent values in Table 3 do not indicate a strong rater effect. Most of the overall convergent correlation values for each PD lie in the middle of the low and high values when each clinician is removed at a time.

Discriminant Validity. We also evaluated the values off the diagonal in Table 2 to index the discriminant validity, or specificity, of the agreement between the self-reported and clinician-rated PD scores. Table 3 summarizes three different types of discriminant validity. First, it presents the traditional discrimination across instruments (i.e., the correlation of SWAP-200 PD scales with nontarget MCMI-III scales). Following this, Table 3 also presents the discriminant correlations *within* instruments (e.g., the correlation of SWAP-200 PD scales with each other), which index the general overlap among scales within each instrument. When calculating the median value for the discriminant validity correlations, we used absolute values to provide a true estimate of the associations with nontarget variables. Without using absolute values for these calculations, the central tendencies can approach zero even when the given values are large (i.e., large positive and large negative values cancel each other out). The minimum and maximum columns, however, provide the actual correlations as the direction of the relationship is informative. The first set of columns

TABLE 2. Multitrait-Multimethod Matrix of SWAP-200 and MCMI-III PD Scales

		SWAP-200										**MCMI-III**								
		PAR	**SZD**	**SZT**	**ATS**	**BDL**	**HST**	**NAR**	**AVD**	**DPD**	**OC**	**PAR**	**SZD**	**SZT**	**ATS**	**BDL**	**HST**	**NAR**	**AVD**	**DPD**
SWAP-200	SZD	**.47**																		
	SZT	**.63**	**.92**																	
	ATS	**.61**	-.16	.08																
	BDL	**.60**	**.32**	**.48**	.22															
	HST	**.27**	-.22	.04	**.55**	**.53**														
	NAR	**.64**	-.12	.04	**.90**	.19	**.54**													
	AVD	.19	**.84**	**.65**	**-.52**	**.32**	**-.33**	**-.44**												
	DPD	-.17	**.48**	**.34**	**-.72**	**.34**	-.10	**-.65**	**.78**											
	OC	.15	**.57**	**.34**	**-.47**	.01	**-.49**	-.26	**.68**	**.44**										
MCMI-III	PAR	-.10	.02	-.01	-.17	-.04	-.03	-.18	.10	.11	.07									
	SZD	.07	**.36**	**.27**	-.20	.13	-.06	-.09	**.39**	.30	.26	**.38**								
	SZT	.06	.08	.05	-.09	.18	-.02	-.02	.16	.11	.13	**.40**	**.47**							
	ATS	.04	-.19	-.03	**.41**	.10	**.33**	.25	**-.34**	**-.35**	**-.42**	-.01	.02	.23						
	BDL	.20	.09	.17	.17	**.40**	.24	.13	.07	.03	-.18	.13	.25	**.53**	**.49**					
	HST	**-.33**	**-.45**	**-.39**	.11	-.11	**.34**	.08	**-.41**	-.14	**-.40**	-.20	**-.51**	**-.31**	.17	-.12				
	NAR	-.12	-.22	-.17	.25	-.08	**.34**	.25	**-.33**	-.17	-.25	-.13	-.22	-.26	.16	-.24	**.62**			
	AVD	.02	**.34**	.23	**-.34**	.04	**-.30**	**-.31**	**.45**	**.32**	.21	**.45**	**.43**	**.53**	-.10	**.31**	**-.58**	**-.59**		
	DPD	-.08	.14	.09	**-.30**	.09	-.19	**-.29**	**.30**	.26	.10	.17	.22	**.48**	-.03	**.51**	**-.31**	**-.55**	**.51**	
	OC	-.25	.15	-.06	**-.48**	**-.28**	**-.32**	**-.33**	**.34**	**.37**	**.43**	.22	.13	.02	**-.76**	**-.52**	-.17	-.06	.21	.00

Notes. $n = 56$. Bolded correlations are significiant at $p < .05$; SWAP-200 = Shedler-Western Assessment Procedure-200; MCMI-III = Millon Clinical Multiaxial Inventory-III; PAR = Paranoid, SZD = Schizoid; SZT = Schizotypal; ATS = Antisocial; BDL = Borderline; HST = Histrionic; NAR = Narcissistic; AVD = Avoidant; DPD = Dependent; OC = Obsessive-Compulsive.

TABLE 3. Convergent and Discriminant Validity of the SWAP-200 and MCMI-III PD Scores

		DISC of SWAP (T scores) with MCMI			DISC within SWAP (T scores)			DISC within MCMI		
	Convergent	*Mdn*	Min	Max	*Mdn*	Min	Max	*Mdn*	Min	Max
Paranoid	–.10	.08	–.33	.20	.47	–.17	.63	.20	–.20	.45
Schizoid	.36	.15	–.45	.34	.47	–.22	.92	.25	–.51	.47
Schizotypal	.05	.17	–.39	.27	.34	–.06	.92	.40	–.31	.53
Antisocial	.40	.20	–.48	.41	.52	–.72	.90	.16	–.76	.49
Borderline	.40	.10	–.28	.19	.32	.01	.60	.24	–.52	.53
Histrionic	.34	.24	–.32	.34	.44	–.49	.55	.31	–.58	.62
Narcissistic	.25	.18	–.33	.13	.44	–.65	.90	.24	–.59	.62
Avoidant	.45	.33	–.41	.39	.52	–.52	.84	.45	–.59	.53
Dependent	.26	.17	–.35	.37	.44	–.72	.78	.31	–.55	.51
OCPD	.43	.21	–.42	.26	.44	–.49	.68	.17	–.76	.22
Median	.35	.17	–.37	.31	.44	–.49	.84	.25	.57	.52
Min	–.10	.08	–.72	–.02	.32	–.72	.55	.16	–.76	.22
Max	.45	.33	–.28	.41	.52	.01	.92	.45	–.20	.76

Note. N = 56; Median discriminant values are calculated based on absolute values. DISC of SWAP (T scores) with MCMI = hetero-method hetero-trait correlations between SWAP-200 scales and MCMI-III scales. DISC within SWAP (T scores) = same-method hetero-trait correlations within SWAP-200 scales.

indicates the discriminant correlation between a given SWAP-200 PD score and all other nontarget MCMI-III PD scores (e.g., SWAP-200 avoidant and MCMI-III narcissistic). The median (absolute) discriminant values across the two instruments ranged from .08 (paranoid) to .33 (avoidant), with a median of .17. These median discriminant values were lower, on an absolute level, than the convergent value for each PD, except for paranoid and schizotypal. However, the final columns within this set indicate that the maximum discriminant values were often as high as, or higher than, the convergent values. For example, SWAP-200 dependent correlated .26 with MCMI-III dependent,

TABLE 4. Highest and Lowest Convergent Validity of the SWAP-200 and MCMI-III PD Scores When Correlations Are Computed Removing One Clinician at a Time

PDs	Lowest Convergent	Highest Convergent
Paranoid	–.19	–.04
Schizoid	.30	.41
Schizotypal	.01	.16
Antisocial	.33	.50
Borderline	.32	.46
Histrionic	.27	.43
Narcissistic	.19	.47
Avoidant	.37	.51
Dependent	.19	.32
OCPD	.34	.49

and .37 with the MCMI-III OCPD score and .32 with MCMI-III avoidant. In sum, five of the 10 SWAP-200 scales (i.e., paranoid, schizotypal, antisocial, histrionic, and dependent) obtained at least one discriminant correlation that was as high as, or higher than, the convergent correlation. Considering these results for discriminant validity *across* the two instruments, we also report the discriminant validity *within* both instruments in Table 3. The median (absolute) discriminant validity correlations for each SWAP-200 scale (i.e., the median correlation of a given SWAP-200 PD scale with all other SWAP-200 PD scales) ranged from .32 to .52, with an overall median of .44. The equivalent discriminant values within the MCMI-III ranged from .16 to .45, with an overall median of .25. Thus, it appeared that the SWAP-200 scales overlapped with each other at a greater rate than the MCMI-III scales did with each other.

DISCUSSION

The present study is the first to compare systematic ratings of all 10 PDs provided by treating clinicians (via the SWAP-200) and their clients (via the MCMI-III). Importantly, this study represents a meaningful increment to the literature for two reasons. First, experienced therapists rating their own clients on the basis of therapeutic contact provided high external validity. Second, the use of two systematic and commonly employed PD instruments provided a realistic estimate of the agreement between clinicians and their clients in practice settings. The dimensional agreement across the 10 PDs ranged from −.10 to .45, with a median of .35. This value is consistent with our hypothesis and comparable to the two prior study of clinician-client agreement using the SWAP (i.e., *Mdn* = .28 from Davidson et al.2003; *Mdn* = .33 from Bradley et al., 2007). The convergent value we reported, along with those from other studies using a version of the SWAP, does appear to be somewhat higher than when clinicians' ratings were collected using brief and unstructured ratings (e.g., *Mdn* = .05 from Chick et al., 1993). Most generally, this reinforces the view that more systematic measurement tools allow for more valid assessments (Samuel et al., 2013). Finally, the convergence we reported for clinicians and clients is remarkably similar to meta-analytic findings regarding the agreement between PD ratings by self-report and other informants (Klonsky et al., 2002).

Nonetheless, another major contribution of our study is the focus on the discriminant validity. This provides an important context because the convergent correlations should not only be sizeable in magnitude but also relatively larger than their absolute correlations with nontarget PDs. In this regard, our findings raise concerns about the specificity of PD ratings on both instruments. Half of the SWAP PD scales correlated as highly with scales from the MCMI-III that assessed PDs other than the target, and half of the MCMI-III scales correlated as highly with nontarget SWAP-200 scales. For example, the SWAP-200 dependent PD scale correlated .26 with the MCMI-III dependent scale, but obtained notable positive correlations with the MCMI-III scales for OCPD (.37), avoidant (.32), and schizoid (.30), as well as a negative correlation with antisocial (−.35). Similarly, the MCMI-III dependent scale obtained correlations with nontarget SWAP-200 scales that were at least as high as the convergent values: avoidant (.30), antisocial (–.30), and narcissistic (–.29). This provides a valuable context for understanding the convergent findings, as it suggests that increased agreement when utilizing systematic PD assessments, such as the SWAP and the MCMI, may simply reflect an increased correlation with all personality pathology measures or reduced measurement error, rather than necessarily improved agreement. In this regard, it might also be possible to hypothesize that the presence of a "general factor" of psychopathology (Rushton, Irwing, & Booth, 2010; but see also Hopwood, Wright, & Donnellan, 2011) and reflection of a nonspecific severity might be accounting for the low discriminant validity yielded by the two target instruments.

Because these correlations are across instruments, this information alone does not necessarily indict the SWAP-200 or the MCMI-III scales as having problematic discriminant validity. Indeed, the detection of elevated correlations among putatively distinct PD constructs is hardly novel (e.g., Lilienfeld, Waldman, & Israel, 1994). No measure should be expected to evince greater discriminant validity than exists within the constructs it measures. Regardless, the most important takeaway point is that the overall level of agreement between clients and clinicians is complicated by levels of discriminant validity that sometimes eclipse convergent values.

Additionally, we note that item overlap across scales on the same measure might have affected discriminant validity. The issue of item overlap within prior versions of the MCMI has been a significant concern, but it was addressed at least partially by the MCMI-III. Nonetheless, there remain a number of items that are scored on more than one PD scale, including some that are scored in opposite directions for different PDs. Such a convention has obvious implications for discriminant validity that would affect our results. Changes for the MCMI-III improved the situation considerably, but this remains an issue in need of further investigation. The issue of item overlap is even more complicated for the SWAP-200 due to the Q-sort item-scoring format. Because the full profile is considered, the score on every item matters for every prototype and subscale. Interestingly, Wood and colleagues (2007) have argued that requiring half of the SWAP items to be rated as a 0 (i.e., *not descriptive*) artificially inflates the discriminant validity of the SWAP-200. Nonetheless, our results do not support this view, or at least suggest that the discriminant validity is no better (and perhaps worse) for the SWAP-200 than for the MCMI-III.

The correlations between the two PD measures considered here are linked to the complex debate about the agreement between self- and clinician-rated measures of PDs more broadly. Thus, we can understand these results at two levels of abstraction. First, as widely illustrated by the literature (Ganellen, 2007; Westen, 1997) one discrepancy between these measures is the obvious contrast in the source of diagnostic information (i.e., clinicians versus clients). However, a second explanation for our findings might also be variability of the convergence across the PDs, which might reflect intrinsic features of the individual PD constructs themselves (e.g., observability, frequency within clinical settings, and level of distress or impairment).

EXPLANATIONS FOR THE OVERALL LEVEL OF AGREEMENT BETWEEN CLIENTS AND CLINICIANS

As has been noted, clinicians (as a type of informant) rely on dramatically different information when making PD judgments than do clients. On the one hand, clients have access to an array of behavior and inner experiences built sequentially across time and multiple contexts, whereas clinicians rely on relatively brief interactions with the clients in a specific context (i.e., typically 1 hour per week in a prescribed one-on-one setting). On the other hand, the two sources also differ tremendously in terms of their training to make such judgments. Clients typically lack any training in the diagnosis of mental disorders, whereas clinicians rely on extensive professional education and experience. Thus, both sources have particular qualities that might recommend them for obtaining information about an individual's personality.

Self-report measures, of course, rely upon the ability of the patient to accurately report on his or her characteristic feelings, behaviors, thoughts, and motives. Nevertheless, the ability of a typical client to do so has been contested (see Ganellen, 2007, for an extensive discussion). Part of self-perception, and so of the experience we have of ourselves, largely resides in implicit memory systems and cannot easily be recalled and translated into words (Berlin, 2011; Roediger, 1990). Furthermore, the access that a person may have to his or her inner psychological world is constantly influenced by a variety of factors, including level of insight and self-awareness, willingness to present one's self objectively, understanding of item content, and mood state variability (Huprich, Bornstein, & Schmitt, 2011). To wit, paranoid PD yielded the lowest agreement in the present study, perhaps due to challenges of self-disclosure. Although these challenges might impair the ability of patients to accurately report such information, there might be similar limits and biases to the accuracy of clinicians' judgments. Morey and Ochoa (1989) demonstrated that clinicians' *own ratings* of individual diagnostic criteria typically did not conform to their holistic diagnoses of the same client. Furthermore, clinicians judgments of others' personality can be affected by their own personality features (Corbitt & Widiger, 1995) and demographic variables of the client (Flanagan & Blashfield, 2003). Taken together, these important differences suggest that neither clinicians' nor clients' ratings should be inherently preferred, but help to understand why their ratings diverge.

EXPLANATIONS FOR DISCREPANCIES ACROSS PERSONALITY DISORDER CONSTRUCTS

It is also worth examining the convergence for individual PD constructs across the two sources. Two central aspects might affect agreement for specific constricts: (a) observability of the features for each PD, and (b) frequency with which they are encountered in clinical settings. Based on the literature on informant reports, it is probable that clinicians and patients agree to a greater extent when the relevant features are more readily observable and/or behaviorally specific (Connelly & Ones, 2010). It is worth noting that four PDs in our study obtained convergent values of .40 or greater: avoidant, obsessive-compulsive, borderline, and antisocial. Antisocial is the most behaviorally specific, whereas avoidant and borderline both heavily involve interpersonal behaviors. This might suggest that therapists, like any other informant, are better equipped to judge those aspects of personality that are most observable (Carlson, Vazire, & Oltmanns, 2013).

In addition, there appears to be a relation between the convergent validity for a given PD and its frequency within clinical settings. In fact, avoidant (14.7%), borderline (9.3%), and obsessive-compulsive (8.7%) are the most prevalent PDs within clinical samples (Zimmerman, Rothschild, & Chelminski, 2005) and had the highest convergence in our sample. It is not clear exactly why frequency might enhance agreement, but it could simply be a property of base rates, such that certain PDs have lower convergence because they occur less frequently. Alternatively, the frequency of each PD in clinical settings might serve to increase the familiarity that clinicians could have in recognizing a PD, resulting in more valid clinician ratings.

LIMITATIONS AND STRENGTHS OF THE PRESENT STUDY

The present study provides the most systematic and comprehensive evaluation of the agreement between clinicians and clients for PD diagnostic ratings. Additionally, our results are strengthened by the high degree of external validity afforded by this naturalistic treatment setting and sizable sample. In this regard, the sample size, although modest compared to studies that rely only on self-report or structured interview ratings, is actually larger than others that have addressed the issue of convergence between diagnoses assigned by practicing therapists with other methods (e.g., *N* = 23 in Davidson et al., 2003; *N* = 54 in Bradley et al., 2007). This reflects the inherent difficulty of obtaining such detailed ratings from busy clinicians and emphasizes the dire need for this type of research in the literature. Finally, a remarkable feature of the present study was the extremely high level of expertise among the clinicians who rated the SWAP-200.

Nonetheless, this effort is not without limitations. Most notable is the temporal lag between the patients' completion of the MCMI-III and the clinicians' ratings on the SWAP-200. This is a complicated issue because the long treatment ensures that the treating clinicians knew the clients particularly well. However, it is obviously quite plausible that patients changed since the beginning of treatment and that these changes decreased the agreement. In essence, the values we report have components of interrater and time-lagged agreement that are difficult to disentangle. That said, we took steps to overcome this limitation by examining agreement across various time frames (including some cases where ratings were completed in close proximity), and there did not appear to be notable patterns of increased or decreased agreement. This may be because PD symptoms remain somewhat stable across time (Samuel et al., 2011). Alternatively, it may implicate more complicated forces that obscure agreement in different, but equal, ways across time. Future work that collects the measures simultaneously from both sources, at multiple points throughout the therapeutic process, would help to dissect the temporal component.

Finally, the present study relied on the traditional conceptualizations of PDs, which have been contested for problems in their validity and excessive comorbidity (Skodol et al., 2011). Despite those concerns, the PDs examined in this study remain the formal definitions in *DSM-5* and thus are routinely used for diagnosis, so providing information that may guide their valid assessment in clinical practice is useful. Nonetheless, future iterations of the *DSM* will likely adopt a dimensional framework that conceptualizes PDs, at least in part, as constellations of maladaptive traits (Skodol, 2014), and so research that examines the clinical application of those traits would be quite valuable (Samuel & Widiger, 2010).

REFERENCES

Barbot, B., Hunter, S. R., Grigorenko, E. L., & Luthar, S. S. (2013). Dynamic of change in pathological personality trait dimensions: A latent change analysis among at-risk women. *Journal of Psychopathology and Behavioral Assessment, 35*(2), 173–185. doi:10.1007/s10862-012-9331-4

Berlin, H. A. (2011). The neural basis of the dynamic unconscious. *Neuropsychoanalysis*, *13*(1), 5–31.

Block, J. (1961). *The Q-sort method in personality assessment and psychiatric research*. Springfield, IL: Charles C Thomas.

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*(2), 225–229. doi: 10.1037/1040-3590.19.2.225

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*(1), 24–31. doi: 10.1097/01.nmd.0000149215.88020.7c

Carlson, E. N., Vazire, S., Oltmanns, T. F. (2013). Self–other knowledge assymetries in personality pathology. *Journal of Personality, 81,* 155–170.

Chick, D., Sheaffer, C. I., Goggin, W. C., & Sison, G. F. (1993). The relationship between MCMI-III personality scales and clinician-generated *DSM-III-R* personality disorder diagnoses. *Journal of Personality Assessment*, *61*(2), 264–276.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159.

Connelly, B. S., & Ones, D. S. (2010). An other perspective on personality: Meta-analytic integration of observers' accuracy and predictive validity. *Psychological Bulletin*, *136*(6), 1092–1122. doi:10.1037/a0021212

Corbitt, E. M., & Widiger, T. A. (1995). Sex differences among the personality disorders: An exploration of the data. *Clinical Psychology: Science and Practice, 2*(3), 225–238.

Davidson, K. M., Obonsawin, M. C., Seils, M., & Patience, L. (2003). Patient and clinician agreement on personality using the SWAP-200. *Journal of Personality Disorders, 17*(3), 208–218. doi:10.1521/pedi.17.3.208.22148

Flanagan, E. H., & Blashfield, R. K. (2003). Gender bias in the diagnosis of personality disorders: The roles of base rates and social stereotypes. *Journal of Personality Disorders, 17*(5), 431–446. doi:10.1521/pedi.17.5.431.22974

Ganellen, R. J. (2007). Assessing normal and abnormal personality functioning: Strengths and weaknesses of self-report, observer, and performance-based methods. *Journal of Personality Assessment*, *89*(1), 30–40.

Gazzillo, F., Lingiardi, V., Peloso, A., Giordani, S., Vesco, S., Zanna, V., . . . Vicari, S. (2013). Personality subtypes in adolescents with anorexia nervosa. *Comprehensive Psychiatry*, *54*(6), 702–712.

Hopwood, C. J., Wright, A. G. C., & Donnellan, M. B. (2011). Evaluating the evidence for the general factor of personality across multiple inventories. *Journal of Research in Personality*, *45*(5), 468–478. doi:10.1016/j.jrp.2011.06.002

Huprich, S. K., Bornstein, R. F., & Schmitt, T. A. (2011). Self-report methodology is insufficient for improving the assessment and classification of Axis II personality disorders. *Journal of Personality Disorders*, *25*(5), *557*–70. doi:10.1521/pedi.2011.25.5.557

Hyler, S. E., Rieder, R. O., Williams, J. B., & Spitzer, R. L. (1989). A comparison of clinical and self-report diagnoses of *DSM-III* personality disorders in 552 patients. *Comprehensive Psychiatry, 30*(2), 170–178. doi:10.1016/0010-440x(89)90070-9

Klein, M. H., Benjamin, L. S., Rosenfeld, R., Treece, C., Husted, J., & Greist, J. H. (1993). The Wisconsin Personality Disorders Inventory:

Development, reliability, and validity. *Journal of Personality Disorders, 7*(4), 285–303.

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 and Practice, 9*(3), 300–311.

Lenzenweger, M. F., Loranger, A. W., Korfine, L., & Neff, C. (1997). Detecting personality disorders in a nonclinical population: Application of a 2-stage procedure for case identification. *Archives of General Psychiatry, 54*(4), 345–351.

Lilienfeld, S. O., Waldman, I. D., & Israel, A. C. (1994). A critical examination of the use of the term and concept of comorbidity in psychopathology research. *Clinical Psychology: Science and Practice*, *1*(1), 71–83.

Lingiardi, V., Shedler, J., & Gazzillo, F. (2006). Assessing personality change in psychotherapy with the SWAP-200: A case study. *Journal of Personality Assessment, 86*(1), 23–32.

Michels, R. (2012). Diagnosing personality disorders. *American Journal of Psychiatry, 169*(3), 241–243.

Millon, T., Davis, R., & Millon, C. (1997). *Millon Clinical Multiaxial Inventory-III manual*. Minneapolis, MN: National Computer Systems.

Morey, L. C. (1991). *Personality Assessment Inventory professional manual*. Odessa, FL: Psychological Assessment Resources.

Morey, L. C., & Ochoa, E. S. (1989). An investigation of adherence to diagnostic criteria: Clinical diagnosis of the *DSM-III* personality disorders. *Journal of Personality Disorders*, *3*(3), 180–192.

Perry, J. C. (1992). Problems and considerations in the valid assessment of personality disorders. *American Journal of Psychiatry, 149*(12), 1645–1653.

Roediger, H. L. (1990). Implicit memory: Retention without remembering. *American Psychologist*, *45*(9), 1043–1056.

Rushton, J. P., Irwing, P., & Booth, T. (2010). A General Factor of Personality (GFP) in the personality disorders: Three studies of the Dimensional Assessment of Personality Pathology-Basic Questionnaire (DAPP-BQ). *Twin Research and Human Genetics*, *13*(4), 301–311.

Samuel, D. B., Hopwood, C. J., Ansell, E. B., Morey, L. C., Sanislow, C. A., Markowitz, J. C., . . . Grilo, C. M. (2011). Comparing the temporal stability of self-reported and interview-based personality disorder. *Journal of Abnormal Psychology, 120,* 670–680.

Samuel, D. B., Sanislow, C. A., Hopwood, C. J., Shea, M. T., Skodol, A. E., Morey, L. C., . . . Grilo, C. M. (2013). Convergent and incremental predictive validity of clinician, self-report, and structured interview diagnoses for personality disorders over 5 years. *Journal of Consulting and Clinical Psychology, 81*(4), 650–659. doi:10.1037/A0032813

Samuel, D. B., & Widiger, T. A. (2010). Comparing personality disorder models: Cross-method assessment of the FFM and *DSM-IV-TR*. *Journal of Personality Disorders*, *24*(6), 721–745. doi:10.1521/pedi.2010.24.6 .721

Skodol, A. E. (2014). Personality disorder classification: Stuck in neutral, how to move forward? *Current Psychiatry Reports*, *16*(10), 480. doi:10.1007/s11920-014-0480-x

Skodol, A. E., Bender, D. S., Morey, L. C., Clark, L. A., Oldham, J. M., Alarcon, R. D., . . . Siever, L. J. (2011). Personality disorder types proposed for *DSM-5*. *Journal of Personality Disorders*, *25*(2), 136–169. doi:10.1521/pedi.2011.25.2.136

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*(7), 895–903.

Westen, D., & Muderrisoglu, S. (2003). Reliability and validity of personality disorder assessment using a systematic clinical interview: Evaluating an alternative to structured interviews. *Journal of Personality Disorders, 17,* 350–368.

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

Westen, D., Shedler, J., Bradley, B., & DeFife, J. A. (2012). An empirically derived taxonomy for personality diagnosis: Bridging science and practice in conceptualizing personality. *American Journal of Psychiatry, 169*, 273–284.

Westen, D., Shedler, J., & Lingiardi, V. (2003). *La valutazione della personalità con la SWAP-200* [The assesment of personality with SWAP-200]. Milan, Italy: Cortina.

Westen, D., & Weinberger, J. (2004). When clinical description becomes statistical prediction. *American Psychologist, 59*(7), 595–613. doi:10.1037/0003-066X.59.7.595

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*(4), 823–836. doi:10.1037/0021-843x.116.4.823

Zennaro, A., Ferracuti, S., Lang, M., Roccaro, G., Roma, P., Sanavio, E., & Horn, S. L. (2013). Diagnostic validity statistics in MCMI-III applied to an Italian sample. *Bollettino di Psicologia Applicata*, *267*, 48–57.

Zimmerman, M., Rothschild, L., & Chelminski, I. (2005). The prevalence of *DSM-IV* personality disorders in psychiatric outpatients. *American Journal of Psychiatry, 162*(10), 1911–1918.