September 1, 2013

Convergent and Incremental Predictive Validity of Clinician, Self-Report, and Structured Interview Diagnoses for Personality Disorders Over 5 Years

Douglas B. Samuel
Charles A. Sanislow, *Wesleyan University*
Christopher J. Hopwood, *Michigan State University*
M. Tracie Shea
Andrew E. Skodol, *University of Arizona*, et al.

Journal of Consulting and Clinical Psychology
2013, Vol. 81, No. 4, 650–659

0022-006X/13/$12.00 DOI: 10.1037/a0032813

Convergent and Incremental Predictive Validity of Clinician, Self-Report, and Structured Interview Diagnoses for Personality Disorders Over 5 Years

Douglas B. Samuel
Yale School of Medicine

Charles A. Sanislow
Wesleyan University

Christopher J. Hopwood
Michigan State University

M. Tracie Shea
Veterans Affairs Medical Center, Providence, Rhode Island, and Alpert Medical School of Brown University

Andrew E. Skodol
The Sunbelt Collaborative, Tucson, Arizona, and University of Arizona College of Medicine

Leslie C. Morey
Texas A&M University

Emily B. Ansell
Yale School of Medicine

John C. Markowitz
New York State Psychiatric Institute, New York, New York, and Columbia University College of Physicians and Surgeons

Mary C. Zanarini
Harvard Medical School

Carlos M. Grilo
Yale School of Medicine

Objective: Research has demonstrated poor agreement between clinician-assigned personality disorder (PD) diagnoses and those generated by self-report questionnaires and semistructured diagnostic interviews. No research has compared prospectively the predictive validity of these methods. We investigated the convergence of these 3 diagnostic methods and tested their relative and incremental validity in predicting independent, multimethod assessments of psychosocial functioning performed prospectively over 5 years. ***Method:*** Participants were 320 patients in the Collaborative Longitudinal Personality Disorders Study diagnosed with PDs by therapist, self-report, and semistructured interview at baseline. We examined the relative incremental validity of therapists' naturalistic ratings relative to these other diagnostic methods for predicting psychosocial functioning at 5-year follow-up. ***Results:*** Hierarchical linear regression analyses revealed that both the self-report questionnaire and semistructured interview PD diagnoses had significant incremental predictive validity over the PD diagnoses assigned by a treating clinician. Although, in some cases, the clinicians' ratings for individual PDs did have validity for predicting subsequent functioning, they did not generally provide incremental prediction beyond the other methods. These findings remained robust in a series of analyses restricted to a subsample of therapist ratings based on clinical contact of 1 year or greater. ***Conclusions:*** These results from a large clinical sample echo previous research documenting limited agreement between clinicians' naturalistic PD diagnoses and those from self-report and semistructured interview methods. They extend prior work by providing the first evidence about the relative predictive validity of these different methods. Our findings challenge the validity of naturalistic PD diagnoses and suggest the use of structured diagnostic instruments.

Keywords: personality disorder, semistructured interview, self-report, diagnostic agreement, clinician

This article was published Online First May 6, 2013.

Douglas B. Samuel, Department of Psychiatry, Yale School of Medicine; Charles A. Sanislow, Department of Psychology, Wesleyan University; Christopher J. Hopwood, Department of Psychology, Michigan State University; M. Tracie Shea, Veterans Affairs Medical Center, Providence, Rhode Island, and Department of Psychiatry, Alpert Medical School of Brown University; Andrew E. Skodol, The Sunbelt Collaborative, Tucson, Arizona, and Department of Psychiatry, University of Arizona College of Medicine; Leslie C. Morey, Department of Psychology, Texas A&M University; Emily B. Ansell, Department of Psychiatry, Yale School of Medicine; John C. Markowitz, New York State Psychiatric Institute, New York, New York, and Department of Psychiatry, Columbia University College of Physicians and Surgeons; Mary C. Zanarini, Department of Psychiatry, Harvard Medical School; Carlos M. Grilo, Department of Psychiatry, Yale School of Medicine.

Writing of this article was supported by the Office of Academic Affiliations, Advanced Fellowship Program in Mental Illness Research and Treatment, Department of Veterans Affairs. Research was supported by National Institute of Mental Health Grants MH 50837, 50838, 50839, 50840, 50850, and MH073708, awared to Charles A. Sanislow. This publication has been reviewed and approved by the Publications Committee of the Collaborative Longitudinal Personality Disorders Study.

Correspondence concerning this article should be addressed to Douglas B. Samuel, who is now at Department of Psychological Sciences, Purdue University, West Lafayette, IN 47907. E-mail: dbsamuel@purdue.edu

The complexity of diagnosing personality disorders (PDs) has been a long-standing issue in psychiatry (Westen, 1997; Zimmerman, 1994). Several methods exist for diagnosing PDs, including semistructured diagnostic interviews, self-report questionnaires, clinician-rated Q-sort instruments, as well as unstructured diagnoses made by treating clinicians (McDermut & Zimmerman, 2005). Although research has relied primarily on semistructured diagnostic interviews and self-report questionnaires, therapists typically base PD diagnoses on their unstructured interviews and clinical contacts with patients (Perry, 1992; Westen, 1997; Zimmerman, 2011). Despite debate regarding the relative merits of different diagnostic methods (Westen & Muderrisoglu, 2003; Zimmerman & Mattia, 1999), no study has yet compared the predictive validity of clinicians' naturalistic PD diagnoses with those from self-report questionnaires or semistructured interviews (Zimmerman, 2011).

Existing research has repeatedly indicated that clinician-generated PD diagnoses do not agree well with those from self-report measures (Davidson, Obonsawin, Seils, & Patience, 2003; Hyler, Rieder, Williams, & Spitzer, 1989; Morey, Blashfield, Webb, & Jewell, 1988; Rossi, Van den Brande, Tobac, Sloore, & Hauben, 2003) or semistructured interviews (Dreessen & Arntz, 1999; Fridell & Hesse, 2006; Samuel & Widiger, 2010). This poor agreement is not unique to PDs, and has been noted for various psychiatric diagnoses (Rettew, Lynch, Achenbach, Dumenci, & Ivanova, 2009). More importantly, fundamental questions regarding the incremental predictive validity of diagnoses assigned by clinicians relative to different methods have not been answered. Research has compared the validity of self- and informant reports of PD (Klein, 2003; Oltmanns & Turkheimer, 2009), but there is a critical need for analogous work comparing clinical diagnoses with other methods. Such work is crucial for determining whether and how different sources of information might be usefully combined. Currently, the optimal approach for how researchers and clinicians *should* most validly identify PDs remains unclear.

Although research on clinical judgment offers reasons for skepticism about the validity of clinician ratings in general (Grove, Zald, Lebow, Snitz, & Nelson, 2000; Meehl, 1954), there are compelling reasons to believe that their PD diagnoses might be useful and valid. Therapists' diagnostic impressions rely on extensive training and take into consideration information about the client's life gleaned across extended periods of clinical interactions. Pilkonis, Heape, Ruddy, and Serrao (1991) noted "clinical judgment, of course, has its own limitations, but it would seem unwise to develop assessment tools that are unrelated to thoughtful clinical experience" (p. 46). In addition, Westen (1997) suggested that clinicians take a holistic approach to diagnosis, situating them well to describe complex personality pathology. Others contend that clinicians' PD ratings are superior to self-report because patients' ability to accurately assess their own personality might be limited by mood states, lack of insight, or presentation biases (Ganellen, 2007; Huprich, Bornstein, & Schmitt, 2011). Finally, Morey et al. (1988) suggested that semistructured diagnostic interviews also might have limitations because "a relatively brief interview situation does not seem particularly well suited to the task of assessing long-term personological characteristics" (p. 47).

Despite these concerns, there are reasons to believe that patient-reported information from semistructured interviews and/or self-report questionnaires can usefully contribute to PD diagnoses (Zimmerman & Mattia, 1999). Thus, although clinicians might not routinely ask direct questions about PD symptoms or use semistructured interviews and self-report questionnaires, they incorporate such information to inform their diagnoses if it is available. Importantly, because semistructured interviews explicitly assess the longitudinal presence of PD symptoms, they might have greater ability to disentangle episodic state artifacts from more durable trait-based PD syndromes (Loranger et al., 1991; Morey et al., 2010).

As treating therapists almost always play the primary role in diagnosing PDs in clinical settings, understanding the relative validity of their impressions carries particular importance. Comparing clinicians' diagnoses with those from self-report questionnaires or semistructured diagnostic interviews would be useful for prospectively predicting clinically relevant outcomes that extend beyond specific diagnostic features, such as psychosocial functioning. We conducted such a comparison using data from the Collaborative Longitudinal Personality Disorders Study (CLPS; Gunderson et al., 2000). The CLPS is well suited for this investigation as the baseline assessment included diagnoses from treating clinicians collected using a modified version of the Personality Assessment Form (PAF; Shea, Glass, Pilkonis, Watkins, & Docherty, 1987; Shea et al., 1990). This allowed them to record the degree to which patients evinced the prototypical characteristics of each of four study PDs (viz., schizotypal, borderline, avoidant, obsessive-compulsive).

The PAF provides a relevant, externally valid method for conducting such an analysis as it closely approximates the way clinicians make PD diagnoses in clinical practice. The PAF's format is also timely, as it uses a prototype-matching approach that mirrors the original proposal for diagnosing PDs in the *Diagnostic and Statistical Manual of Mental Disorders*, fifth edition (*DSM–5*). In fact, the PAF and research that had used it were cited as primary support for the Work Group's proposal (Skodol, Bender, et al., 2011; Skodol, Clark, et al., 2011). This proposal subsequently was criticized by a number of PD scholars (Pilkonis, Hallquist, Morse, & Stepp, 2011; Widiger, 2011; Zimmerman, 2011) and abandoned. Nonetheless, other prominent researchers and clinicians have strongly argued that the prototype-matching approach should become the standard method of PD diagnosis (Shedler et al., 2010).

The benefit and goal of using the PAF for collecting clinicians' impressions is to maximize *external validity* (i.e., most closely match the type of PD diagnoses typically made in clinical practice), not to provide equivalence with other methods (Westen & Weinberger, 2004). Westen and colleagues have demonstrated that when clinicians administer a systematic clinical interview (i.e., the Clinical Diagnostic Interview, CDI; Westen, 2004) and record their impressions using the Shedler-Westen Assessment Procedure (SWAP; Westen & Shedler, 1999), their PD diagnoses become more reliable across independent raters (Westen & Muderrisoglu, 2006; Westen, Shedler, Bradley, & DeFife, 2012). Although informative, such a diagnostic strategy (i.e., a 2-hr administration of the CDI followed by the sorting of 200 SWAP items) is not standard practice in naturalistic settings. Perhaps recognizing this, Westen and his colleagues have also been the primary proponents of the prototype-matching approach (Shedler & Westen, 2004; Westen, DeFife, Bradley, & Hilsenroth, 2010; Westen, Shedler, & Bradley, 2006) that helped inform the original *DSM–5* proposal (Skodol, Bender, et al., 2011; Skodol, Clark, et al., 2011). The

PAF's prototype-matching format makes it a reasonable choice for collecting treating clinicians' PD diagnoses in this study.

We compared the incremental validity of clinicians' diagnoses of these four PDs assigned via the PAF with those generated by a semistructured interview and self-report questionnaire for predicting psychosocial functioning assessed prospectively over 5 years. Given the published support for the validity of the prototype-matching approach (Westen et al., 2012), we hypothesized that clinicians' PAF ratings would account for variance in functioning beyond that captured by self-report questionnaires or semistructured interviews. Nonetheless, we also recognized that all previous findings concerning the relative validity of alternative diagnostic methods have suggested that the methods are mutually informative (Hopwood et al., 2008; Klein, 2003). Thus, we also hypothesized that the self-report and semistructured interview methods would have unique strengths and demonstrate incremental predictive validity beyond the clinician-assigned diagnoses. Finally, to account for inadequate familiarity with patients that might disadvantage the clinicians' PAF ratings, we conducted additional analyses using only the subset of cases, whom clinicians had treated for at least 1 year prior to providing the diagnoses. This choice of a 1-year interval of treatment ensured adequate familiarity with a patient's personality pathology.

Method

Study participants were drawn from the 668 participants recruited from the multiple CLPS clinical sites. Appropriate Institutional Review Boards approved the study. Participants who provided written, informed consent underwent diagnostic interviews and completed self-report questionnaires as part of a standardized battery. Detailed recruitment and diagnostic procedures have been published elsewhere (Gunderson et al., 2000). Briefly, participants were assigned to one of four PD groups (borderline, avoidant, schizotypal, and obsessive-compulsive [OC]), or to major depressive disorder (MDD) without any PD. These PD diagnostic assignments were based on the Diagnostic Interview for *DSM–IV* Personality Disorders (DIPD-IV; Zanarini, Frankenburg, Sickel, & Yong, 1996), reliably administered by trained research personnel. For inclusion, these diagnoses required confirmation by a self-report questionnaire (e.g., Schedule for Nonadaptive and Adaptive Personality–2; SNAP-2; Clark, Simms, Wu, & Casillas, in press) and/or the treating clinician's PAF ratings. Furthermore, because inclusion demanded either a self-report or clinician-assigned diagnosis, in a subset of participants the semistructured interview-assigned diagnosis disagreed with the clinicians' ratings and was instead confirmed by the self-report questionnaire.

Participants used for the current analyses were 320 individuals from the CLPS with available PAF ratings completed by a treating clinician at baseline. Independent sample *t* tests and chi-square tests demonstrated no significant differences between participants with PAF scores and the larger CLPS sample in gender, age, or ethnicity. Independent samples *t* tests revealed that this subsample differed in diagnosis and functioning, perhaps reflecting that participants with PAF ratings were in ongoing psychiatric or psychological treatment. Participants with PAF ratings met more criteria for borderline PD according to the DIPD-IV at baseline ($M = 4.4$, $SD = 2.7$) than did those without available PAF ratings ($M = 2.6$, $SD = 2.5$), $t(729) = 9.3$, $p < .01$. Differences for the other three studied PDs on the DIPD-IV were nonsignificant. Baseline SNAP-2 PD scores were significantly greater for the studied group for all four PDs. Participants with available PAF ratings did not differ from those without in terms of psychosocial functioning measured by the Social Adjustment Scale, Self-Report (SAS-SR), $t(700) = 1.1$, $p = .28$, but did differ significantly according to the Longitudinal Interval Follow-Up Evaluation (LIFE), $t(727) = 5.3$, $p < .01$.

Average age of the participants at baseline was 32.9 years ($SD = 7.9$, range = 18–45); 199 (62%) were women; the ethnic breakdown was 237 (74%) Caucasian, 35 (11%) African American, 39 (12%) Hispanic, six (2%) Asian American, and three (1%) "other." Of the participants, 73 (23%) were assigned to the avoidant, 128 (40%) to the borderline, 54 (17%) to the obsessive-compulsive, 37 (12%) to the schizotypal, and 28 (9%) to the MDD without PD groups. Clinicians reported clinical contact with the patients ranging from 0 to 884 weeks, with a mean of 53.7 ($SD = 89.7$) at the time of providing the PAF ratings. Their confidence in their diagnostic ratings evinced a mean of 2.26 (on a 1– 5 metric, where 1 = *high* and 5 = *low*; $SD = 1.12$).

Personality Disorder Measures

DIPD-IV (Zanarini et al., 1996). The DIPD-IV is a semistructured diagnostic interview for assessing the *Diagnostic and Statistical Manual of Mental Disorders*, fourth edition (*DSM–IV*; American Psychiatric Association, 1994) PDs. Each criterion is assessed with one or more questions rated on a 3-point scale (0 = *not present*; 1 = *present but of uncertain clinical significance*; 2 = *present and clinically significant*). The DIPD-IV requires that criteria be pervasive, present for at least 2 years, and characteristic of the person for most of his or her adult life. In the CLPS sample, interrater reliability (based on 84 pairs of raters) kappa coefficients ranged from .58 to 1.00 (Zanarini et al., 2000). The current report considered only the DIPD-IV scores for the four PDs studied in CLPS.

SNAP-2 (Clark et al., in press). Comprising 390 true/false statements, the SNAP-2 provides a self-report assessment of 12 pathological personality traits derived from an iterative factor analytic process. The SNAP-2 includes scales assessing the *DSM–IV* PDs, ranging in length from 19 (avoidant) to 34 (antisocial) items. Although most *DSM–IV* PD scale items are also scored for one of the trait scales, a number of items were added to explicitly tap additional content. The PD scales can be scored dimensionally or by individual diagnostic criteria to yield categorical diagnoses. In the full CLPS sample, the SNAP-2 PD scale internal consistencies ranged from .69 (OCPD) to .88 (avoidant), with an overall median of .83. The SNAP-2 PD scores correlate consistently with those from other self-report PD inventories (Widiger & Boyd, 2009) and structured PD diagnostic interviews (Samuel et al., 2011). The current report only included the SNAP-2 scores for the four CLPS PDs.

PAF (Shea et al., 1987, 1990). The PAF was adapted for the *DSM–IV* PDs from a measure developed for the National Institute of Mental Health Treatment of Depression Collaborative Research Program (Elkin, Parloff, Hadley, & Autry, 1985). Its purpose was to provide a standardized method to quantify clinicians' routine clinical diagnoses. Thus, it was designed to maximize external validity and mirror the type of PD ratings and diagnoses made in

clinical practice. The PAF used in CLPS contained three to four sentence prototypical descriptions for each of the four PDs studied (schizotypal, borderline, avoidant, and obsessive-compulsive) as well as several "cues" to aid clinicians in rating a patient's match to the prototypes. The instrument is available by request to the first author. Clinicians rated all four of the studied PDs on a 1–6 scale, where 1 indicated *not at all* and 6 indicated that the patient matched the prototype *to an extreme degree*. Consistent with previous research (Shea et al., 1990), a score ≥ 4 indicated a categorical diagnosis. Clinicians could also indicate *no information or insufficient data* for a particular PD, although they used this only rarely (24 times across the four PDs in the sample of 320). Those values were recoded as missing for the current analyses. The mean PAF ratings were 1.95 (SD = 1.20) for schizotypal PD, 2.94 (SD = 1.55) for borderline PD, 2.49 (SD = 1.34) for avoidant PD, and 2.08 (SD = 1.33) for OCPD.

Psychosocial Functioning Measures Serving as Independent External Criteria

Multiple measures of psychosocial functioning served as external outcome criteria. These were independent of specific PD symptoms and used two independent assessment methods. Both aspects are crucial for the current purposes, as independent, external criteria provide the only opportunity to discriminate validity among different methods of PD diagnosis. To assess psychosocial functioning, CLPS research team interviewers administered the LIFE (Keller et al., 1987), a structured interview assessing functioning in interpersonal relationships and occupational and recreational domains. Most areas of functioning are rated on 5-point severity scales (1 = *no impairment, high level of functioning or very good functioning* and 5 = *severe impairment or very poor functioning*). Participants also completed the SAS-SR (Weissman & Bothwell, 1976), a self-report instrument yielding estimates of interpersonal, occupational, and recreational functioning. The LIFE and SAS-SR were administered at baseline and repeated at predetermined intervals, including the 5-year follow-up. The same interviewers administered both interviews (i.e., the LIFE and DIPD-IV) at a given assessment interval; however, it was unlikely that the interviewer who administered the DIPD-IV at baseline also administered the LIFE at 5-year follow-up.

Data Analytic Procedures

We first examined the convergent validity of clinicians' PAF diagnoses with those from a semistructured diagnostic interview (DIPD-IV) and self-report questionnaire (SNAP-2). PAF dimensional ratings were compared with those from the DIPD-IV and SNAP-2 (all at baseline) for their ability to predict functional outcomes at the 60-month follow-up (via the LIFE and SAS-SR) using hierarchical regression analyses. For example, the clinicians' baseline PAF ratings for the four PDs were entered simultaneously in one step, followed by the baseline PDs ratings from the SNAP-2. This was then repeated with the order of entry reversed. To account for possible contamination due to shared method variance, we conducted these analyses separately using the self-report criterion and again with the interview-based criterion variable.

PAF diagnoses had been used to confirm the DIPD-IV diagnosis for a subset of participants, creating a potential confound. Although our use of functional outcomes rather than diagnostic information as criteria attenuates this possibility, we nonetheless examined it by performing a parallel set of analyses restricted to a subsample of 110 participants for whom the PAF disagreed with the DIPD-IV at baseline and thus was not required for study inclusion. In this subsample, PAF ratings would potentially have greater ability to increment the DIPD-IV scores.

Results

Categorical and Dimensional Agreement

Table 1 provides the agreement between PAF ratings and those from the DIPD-IV and SNAP-2. Categorical agreement (kappas) between treating clinicians' diagnoses and the semistructured diagnostic interview ranged from of .21 (avoidant) to .42 (schizotypal), whereas dimensional agreement (Pearson correlations) ranged from .30 (avoidant) to .44 (borderline). Agreement between clinicians' ratings and self-report questionnaire was lower than between clinicians' ratings and semistructured diagnostic interviews, with kappas ranging from .00 (OCPD) to .20 (borderline) and Pearson correlations ranging from .18 (schizotypal) to .28 (borderline). For context, we note that agreement between DIPD-IV and SNAP-2 in the current sample ranged from .25 (OCPD) to .51 (avoidant) for categorical diagnoses and from .57 (schizotypal) to .72 (avoidant) for dimensional ratings.

Incremental Predictive Validity

Tables 2–6 summarize the hierarchical regression analyses. Table 2 shows that the DIPD-IV provided significant increment beyond the PAF for predicting functioning assessed by both the SAS-SR and LIFE. In contrast, clinicians' ratings did not significantly increment the DIPD-IV interview results for either criterion. The nonsignificant ΔR^2 when the PAF block was added does not indicate that all PAF diagnoses lacked validity, as the individual schizotypal rating from the PAF was a significant predictor (β = .15; $p < .05$). Table 3 summarizes the parallel series of analyses on

Table 1
Dimensional and Categorical Agreement of Clinician PD Diagnostic Ratings With Interview Generated and Self-Report PD Scores

	DIPD-IV criteria counts		SNAP-2 PD scores	
PAF ratings	κ	*r*	κ	*r*
Schizotypal	.42	.40	.01	.18
Borderline	.38	.44	.20	.28
Avoidant	.21	.30	.14	.23
OCPD	.24	.30	.00	.20

Note. n = 320. Kappa between diagnoses provided by PAF (≥4) and from DIPD-IV and SNAP-2 (meeting diagnostic criteria threshold). Dimensional agreements represent Pearson correlations of PAF ratings (1–6) with scores from DIPD-IV and SNAP-2. PD = personality disorder; PAF = Personality Assessment Form; DIPD-IV = Diagnostic Interview for *DSM–IV* Personality Disorders; SNAP-2 = Schedule for Nonadaptive and Adaptive Personality–2; OCPD = obsessive-compulsive personality disorder.

Table 2
Hierarchical Multiple Regression Analyses Comparing Baseline Clinician and Semistructured Interview PD Ratings for Predicting Psychosocial Functioning at 60 Months

	Source of psychosocial functioning rating			
	SAS-SR		LIFE	
Predictor	ΔR^2	β	ΔR^2	β
Step 1	.06*		.10***	
PAF Schizotypal		.17*		.25***
PAF Borderline		.15*		.17**
PAF Avoidant		.02		.06
PAF OCPD		−.01		−.10
Step 2	.11***		.09***	
DIPD-IV Schizotypal		.19*		.19**
DIPD-IV Borderline		.15		.07
DIPD-IV Avoidant		.17*		.15*
DIPD-IV OCPD		.00		.05
Total R^2	.17***		.19***	
n	193		234	
Step 1	.15***		.16***	
DIPD-IV Schizotypal		.22**		.26***
DIPD-IV Borderline		.15*		.12
DIPD-IV Avoidant		.16*		.15*
DIPD-IV OCPD		.02		.02
Step 2	.01		.03	
PAF Schizotypal		.09		.15*
PAF Borderline		.05		.11
PAF Avoidant		−.04		.02
PAF OCPD		.06		−.08
Total R^2	.17***		.19***	
n	193		234	

Note. Outcome variables are means of interpersonal, recreation, and work functioning assessed by the SAS-SR and LIFE, respectively. PD = personality disorder; SAS-SR = Social Adjustment Scale, Self-Report; LIFE = Longitudinal Interval Follow-Up Evaluation; OCPD = obsessive-compulsive personality disorder; PAF = Dimensional (1–6) ratings on the Personality Assessment Form; DIPD-IV = criterion count from Diagnostic Interview for *DSM–IV* Personality Disorders.
* $p < .05$. ** $p < .01$. *** $p < .001$.

a subsample of participants whose study inclusion was not confirmed by the PAF. The results were nearly identical with the SAS-SR as criterion: The DIPD-IV provided increment over the PAF, but not vice versa. In contrast with the findings in Table 2, neither instrument incremented the other in predicting the LIFE in the subsample for which the PAF and DIPD-IV disagreed at baseline.

Table 4 summarizes regression analyses comparing SNAP-2 ratings with the PAF in predicting functioning assessed by the SAS-SR and LIFE. The SNAP-2 significantly incremented validity over the clinicians' PAF ratings when using either criterion measure. Although the PAF failed to increment the SNAP-2 for predicting the SAS-SR composite, the PAF did provide significant increment when the LIFE was the criterion.

Does Increased Clinician Familiarity Improve Diagnostic Ratings?

To determine whether clinicians' familiarity with their patients influenced findings, we repeated these analyses using a subsample of clinicians who treated patients for more than 1 year. Ninety clinicians had this level of familiarity, and 60-month follow-up data were available for 73 using the LIFE and 62 using the SAS-SR.

Regression analyses restricted to this subsample revealed findings similar to the overall study group. Table 5 compares the DIPD-IV and PAF in this subsample of clinicians with extensive familiarity. Using the LIFE as criterion, the DIPD-IV significantly incremented PAF ratings, but again the PAF added no significant predictive validity to the DIPD-IV. Using the SAS-SR as criterion, the DIPD-IV's increment of the PAF fell just short of statistical significance, whereas the PAF failed to increment the DIPD-IV appreciably. Table 6 summarizes the regression analyses, comparing the SNAP-2 and the PAF in the subsample, and indicates that the SNAP-2 significantly incremented the PAF but that the reverse did not occur when using either psychosocial functioning measure. These results suggest that clinicians' naturalistic PD diagnoses, relative to more structured methods, are not more valid even after a substantial period of treatment interaction.

Table 3
Hierarchical Multiple Regression Analyses Comparing Baseline Clinician and Semistructured Interview PD Ratings for Predicting Psychosocial Functioning at 60 Months Using Only Cases Where PAF Did Not Confirm DIPD-IV for Study Inclusion

	Source of psychosocial functioning rating			
	SAS-SR		LIFE	
Predictor	ΔR^2	β	ΔR^2	β
Step 1	.07		.13**	
PAF Schizotypal		.08		.16
PAF Borderline		.25*		.32**
PAF Avoidant		−.03		.10
PAF OCPD		−.03		−.16
Step 2	.11*		.05	
DIPD-IV Schizotypal		.24*		.15
DIPD-IV Borderline		.17		.09
DIPD-IV Avoidant		.04		.06
DIPD-IV OCPD		−.03		.06
Total R^2	.18***		.19***	
n	99		117	
Step 1	.15**		.12**	
DIPD-IV Schizotypal		.23*		.21*
DIPD-IV Borderline		.19		.14
DIPD-IV Avoidant		.08		.09
DIPD-IV OCPD		−.03		.05
Step 2	.03		.06	
PAF Schizotypal		.01		.12
PAF Borderline		.14		.24*
PAF Avoidant		−.06		.08
PAF OCPD		.06		−.13
Total R^2	.18***		.19***	
n	99		117	

Note. Outcome variables are means of interpersonal, recreation, and work functioning assessed by the SAS-SR and LIFE, respectively. PD = personality disorder; SAS-SR = Social Adjustment Scale, Self-Report; LIFE = Longitudinal Interval Follow-Up Evaluation; OCPD = obsessive-compulsive personality disorder; PAF = Dimensional (1–6) ratings on the Personality Assessment Form; DIPD-IV = criterion count from the Diagnostic Interview for *DSM–IV* Personality Disorders.
* $p < .05$. ** $p < .01$. *** $p < .001$.

Table 4
Hierarchical Multiple Regression Analyses Comparing Baseline Clinician and Patient-Reported PD Ratings for Predicting Psychosocial Functioning at 60 Months

	Source of psychosocial functioning rating			
	SAS-SR		LIFE	
Predictor	ΔR^2	β	ΔR^2	β
Step 1	.08*		.13***	
PAF Schizotypal		.21**		.29***
PAF Borderline		.18*		.19**
PAF Avoidant		.01		.06
PAF OCPD		−.01		−.14
Step 2	.13***		.05*	
SNAP-2 Schizotypal		.14		.15
SNAP-2 Borderline		.18		−.02
SNAP-2 Avoidant		.16		.13
SNAP-2 OCPD		−.10		.00
Total R^2	.21***		.18***	
n	161		194	
Step 1	.18***		.10***	
SNAP-2 Schizotypal		.17		.19
SNAP-2 Borderline		.19		.05
SNAP-2 Avoidant		.16		.14
SNAP-2 OCPD		−.09		−.04
Step 2	.03		.08**	
PAF Schizotypal		.14		.24**
PAF Borderline		.09		.15*
PAF Avoidant		−.03		.02
PAF OCPD		.07		−.13
Total R^2	.21***		.18***	
n	161			

Note. Outcome variables are means of interpersonal, recreation, and work functioning assessed by the SAS-SR and LIFE, respectively. PD = personality disorder; SAS-SR = Social Adjustment Scale, Self-Report; LIFE = Longitudinal Interval Follow-Up Evaluation; SNAP-2 = Schedule for Nonadaptive and Adaptive Personality–2; OCPD = obsessive-compulsive personality disorder; PAF = Dimensional (1–6) ratings on Personality Assessment Form; SNAP-2 = sum of items from the Schedule for Nonadaptive and Adaptive Personality–2 PD scales.
* $p < .05$. ** $p < .01$. *** $p < .001$.

Discussion

In the current study, we explicitly compared the value of PD diagnoses provided by clinicians via the PAF with those from a semistructured interview and self-report questionnaire for prospectively predicting psychosocial functioning in a large clinical sample. The primary and novel findings were that clinicians' diagnostic ratings were collectively never more informative than those from a semistructured diagnostic interview and only provided significant incremental predictive validity beyond self-report ratings in one of four comparisons. In contrast, semistructured interview and self-report questionnaire PD diagnoses consistently (in eight out of 10 comparisons) predicted significant variance in psychosocial functioning beyond clinician ratings.

These findings were robust despite our efforts, against the tide of experimentwise error, to restrict analyses to subsamples that one would expect to enhance the validity of clinician ratings. Notably, clinicians' diagnoses did not increment the other methods even when the clinician had treated the patient for over 1 year. These findings concern the simultaneous entry of the four study diagnoses and do not suggest that clinicians' individual diagnoses were always devoid of predictive utility. When the LIFE was used as the criterion, the PAF schizotypal and borderline ratings emerged as significant predictors in some analyses, suggesting these diagnostic ratings were unique predictors of subsequent functioning. Thus, our findings suggest that clinicians' prototype ratings on the PAF *do* have some validity but that when considered collectively, they simply have less validity than those from a self-report questionnaire or semistructured interview for predicting functioning after 5 years. This finding raises questions about the validity of PD diagnoses provided by therapists in routine clinical practice and reduces confidence in the ability of the prototype-matching approach to successfully remedy this concern.

Our findings replicate and extend the few available studies documenting that alternative methods of assessing PDs demonstrate incremental predictive validity relative to one another. Semistructured diagnostic interviews (Hopwood et al., 2008) and informant reports (Klein, 2003) have been found to provide incremental predictive validity beyond self-reports. In those studies, however, self-reports also incremented the other methods. Our novel finding is that this was typically not true for treating clinicians' naturalistic

Table 5
Hierarchical Multiple Regression Analyses Comparing Baseline Clinician and Semistructured Interview PD Ratings for Predicting Psychosocial Functioning at 60 Months Using Only Clinicians Who Had Treated the Patient 1 Year or More

	Source of psychosocial functioning rating			
	SAS-SR		LIFE	
Predictor	ΔR^2	β	ΔR^2	β
Step 1	.08		.12	
PAF Schizotypal		.15		.25*
PAF Borderline		.19		.23
PAF Avoidant		.12		.05
PAF OCPD		−.05		−.15
Step 2	.14		.17*	
DIPD-IV Schizotypal		.11		.17
DIPD-IV Borderline		.14		.22
DIPD-IV Avoidant		.30*		.26*
DIPD-IV OCPD		.00		.07
Total R^2	.21**		.29***	
n	58		69	
Step 1	.19*		.23**	
DIPD-IV Schizotypal		.18		.27*
DIPD-IV Borderline		.17		.25*
DIPD-IV Avoidant		.28*		.20
DIPD-IV OCPD		−.03		.00
Step 2	.03		.06	
PAF Schizotypal		.14		.21
PAF Borderline		.11		.10
PAF Avoidant		−.01		−.06
PAF OCPD		−.04		−.16
Total R^2	.21**		.29***	
n	58		69	

Note. Outcome variables are means of interpersonal, recreation, and work functioning assessed by the SAS-SR and LIFE, respectively. PD = personality disorder; SAS-SR = Social Adjustment Scale, Self-Report; LIFE = Longitudinal Interval Follow-Up Evaluation; OCPD = obsessive-compulsive personality disorder; PAF = Dimensional (1–6) ratings on the Personality Assessment Form; DIPD-IV = criterion count from the Diagnostic Interview for *DSM–IV* Personality Disorders.
* $p < .05$. ** $p < .01$. *** $p < .001$.

Table 6
Hierarchical Multiple Regression Analyses Comparing Baseline Clinician and Patient-Reported PD Ratings for Predicting Psychosocial Functioning at 60 Months Using Only Clinicians Who Had Treated the Patient 1 Year or More

	Source of psychosocial functioning rating			
	SAS-SR		LIFE	
Predictor	ΔR^2	β	ΔR^2	β
Step 1	.08		.14	
PAF Schizotypal		.12		.29*
PAF Borderline		.20		.25
PAF Avoidant		.13		.06
PAF OCPD		.04		−.13
Step 2	.24*		.19*	
SNAP-2 Schizotypal		.05		.18
SNAP-2 Borderline		.31		.06
SNAP-2 Avoidant		.24		.28
SNAP-2 OCPD		−.09		−.04
Total R^2	.32***		.34***	
n	49		57	
Step 1	.28**		.24**	
SNAP-2 Schizotypal		.08		.16
SNAP-2 Borderline		.28		.16
SNAP-2 Avoidant		.28		.27
SNAP-2 OCPD		−.10		−.07
Step 2	.04		.10	
PAF Schizotypal		.09		.27*
PAF Borderline		.14		.21
PAF Avoidant		.05		−.05
PAF OCPD		.10		−.12
Total R^2	.32***		.34***	
n				

Note. Outcome variables are means of interpersonal, recreation, and work functioning assessed by the SAS-SR and LIFE, respectively. PD = personality disorder; SAS-SR = Social Adjustment Scale, Self-Report; LIFE = Longitudinal Interval Follow-Up Evaluation; OCPD = obsessive-compulsive personality disorder; PAF = Dimensional (1–6) ratings on the Personality Assessment Form; SNAP-2 = sum of items from the Schedule for Nonadaptive and Adaptive Personality–2 PD scales.
* $p < .05$. ** $p < .01$. *** $p < .001$.

diagnoses of PDs, which yielded little predictive utility beyond interview and self-report methods. This suggests that the assignment of PD diagnoses in clinical practice might be more valid when informed by data from a semistructured interview or a self-report questionnaire.

Our findings regarding the degree of convergence across diagnostic methods again echo previous research. Clinician-assigned PD diagnoses show only modest agreement with semistructured diagnostic interviews and still less agreement with self-report questionnaires (Hyler et al., 1989). In fact, only one kappa between the DIPD-IV and the PAF ($K = .42$ for schizotypal) would even qualify as "fair" (e.g., 0.4–0.6), per Cicchetti (1994). Although these cross-method agreements were poor in an absolute sense, they actually exceeded those obtained in previous research (Hyler et al., 1989; Morey et al., 1988; Samuel & Widiger, 2010), and might have been inflated by the use of PAF ratings to confirm DIPD-IV diagnoses for some study inclusion decisions. Nonetheless, our findings suggest that clinicians' naturalistic PD diagnoses, even when recorded in a standardized format such as the PAF, do not agree well with other methods. This disconnect limits the potential for evidence-based practice because research on PDs, typically based on diagnostic interviews, might not generalize to clinical practice.

It is perhaps unsurprising that clinician diagnoses and interviews/questionnaires yield different PD estimates, as each method approaches the task differently. For example, Westen (1997) reported that practicing clinicians rely primarily on patients' narratives and behaviors in the consulting room when assigning PD diagnoses, but only rarely use explicit questions about *DSM–IV* symptoms, which are the hallmark of questionnaires and diagnostic interviews. Although some might see this approach, which allows clinicians to freely follow their intuition unconstrained by specific criteria or symptoms, as a strength, others contend it dilutes the reliability and validity of psychiatric diagnoses (Zimmerman, 2011).

One plausible explanation for our findings regarding the limited predictive validity of clinicians' naturalistic PD diagnoses is that therapists imperfectly collect and organize information obtained during clinical interactions (Grove et al., 2000; Meehl, 1954). Errors could occur during the diagnostic interview: A clinician might ask idiosyncratic questions and neglect the full array of diagnostic criteria (Westen & Weinberger, 2004). It might also reflect the fact that clinicians use observed behaviors to inform their diagnoses, yet typically interact with their patients in only a single setting (i.e., the consulting room), which has proscribed social roles that might restrict patients' behavioral repertoires. Even if clinicians obtain all relevant information, cognitive biases may enter during transcription and encoding. For example, research has demonstrated that salient features (e.g., self-harm for borderline PD) are more heavily weighted than others and often lead to misdiagnosis (Blashfield & Herkov, 1996; Morey & Ochoa, 1989).

An additional possibility is that our results reflect the *method* of aggregating data (i.e., the instruments) as much as the *source* (i.e., clinician vs. patient). Westen and Weinberger (2004) argued that clinical judgment is often conflated with the nonstandardized aggregation of data, which can allow bias and hamper validity. Thus, our results could suggest that the PAF, which relies on global impressions of prototypes, introduces error into the diagnostic process. In contrast, more systematic assessments by self-report questionnaires and semistructured diagnostic interviews might limit this possibility. Research supports this view in demonstrating that clinicians' global impressions often converge poorly with *their own* systematic diagnostic ratings of the same patient (Morey & Ochoa, 1989). Although clinicians might prefer to diagnose PDs in terms of a holistic match to categorical prototypes (Rottman, Ahn, Sanislow, & Kim, 2009), this method is prone to reasoning errors that limit validity (Zimmerman, 2011).

Future research that examines the incremental predictive validity of clinicians' diagnoses derived from more structured assessments, such as therapists completing the SWAP (Westen & Shedler, 1999), the Personality Disorder Schedule (Nestadt et al., 2012), or even an informant version of an existing PD questionnaire, would help to address this possibility. Nonetheless, it is important to recognize that the PAF is more naturalistic to clinical practice than having clinicians describe patients using, for example, the SWAP. Thus, data supporting such a hypothesis would still recommend a shift in the prevailing diagnostic procedures.

Clinical Implications

A primary implication of the current findings for clinical practice is that the use of semistructured diagnostic interviews and/or self-report questionnaires would improve the validity of PD diagnoses in clinical practice. Although the validity of methods for aggregating clinicians' descriptions might vary, the current results disfavor clinical applications of the PAF specifically, and the prototype-matching technique more generally. This finding is timely and practical, considering the *DSM–5* Personality and Personality Disorders Work Group's initial proposal of a prototype-matching technique for diagnosing PDs as well as continuing advocacy for that method (Shedler et al., 2010). Our findings indicate the relatively lower validity of a prototype-matching approach and support the recent decision to abandon it for the *DSM–5*. Rather, our results suggest that clinicians use standardized assessment instruments to inform PD diagnoses.

Our results temper concerns about the limitations of self-report for assessing personality pathology (Huprich et al., 2011). Although it is reasonable to consider possible response sets that might influence results, self-report questionnaires have advantages over other methods, including inexpensive data collection and available community and clinical norms. Furthermore, the individual who completes them is intimately familiar with his or her own thoughts, feelings, and behaviors over an entire lifetime. That the SNAP-2 PD scores predicted functional outcomes better than clinicians' PD ratings provides important evidence supporting the utility of self-report instruments for assessing personality pathology. Although any individual's description of a person (whether rating oneself or a patient treated over many years) may contain biases, this appeared less problematic for self-report than for clinician ratings in the current study.

A final implication is that our findings regarding the relative validity of clinicians' routine, unstructured diagnoses might extrapolate beyond PDs. We obtained kappa values comparable to those for most other psychiatric disorders (Rettew et al., 2009), leaving little reason to believe our results are peculiar to PDs. Future research exploring the validity of other psychiatric diagnoses provided by clinicians in routine practice warrants attention.

Limitations

In the current study, we examined a large, carefully diagnosed clinical sample with well-validated criterion measures to provide the first data on the relative validity of clinicians' PD diagnoses for predicting prospective psychosocial functioning. Although this sample is well suited for addressing such a question, this was not the original aim of the data collection. Participants entered the study only if they were diagnosed with a study PD by DIPD-IV *and* confirmed by another method (PAF and/or a self-report questionnaire). This sampling strategy excluded other potential participants relevant to the current analyses, such as individuals diagnosed with a PD by the PAF but not according to the DIPD-IV. This strategy possibly enhanced the validity of the DIPD-IV, as some alternative method always buttressed its diagnoses. This limitation does not apply to the SNAP-2 and PAF comparison, which were entirely independent from each other. Clearly though, these findings need replication and extension in additional samples.

Although the PAF successfully approximated both the naturalistic diagnosis of PDs and the prototype-matching system originally proposed for *DSM–5*, it has limitations. It did not allow collection of the therapists' demographic and training information, which would have been helpful in investigating the findings. Future research should use more structured instruments for collecting therapist ratings, thereby controlling the method of aggregation (Westen & Weinberger, 2004) and more directly focusing on the relative merits of the source (e.g., clinician vs. self-report). This could be a clinician-specific instrument such as the SWAP-II, but it would be informative to have clinicians complete an informant version of a self-report measure such as the Personality Instrument for *DSM–5* (Krueger, Derringer, Markon, Watson, & Skodol, 2012).

A major strength of the study was the use of two measures of psychosocial functioning as criterion variables, rather than solely relying on diagnostic outcomes or convergence. Nonetheless, it would have been ideal to collect clinician ratings of functioning that could have been used as another criterion. We note, in this regard, that the lone finding where the PAF provided incremental validity over another diagnostic method was the comparison with the SNAP-2 using the LIFE, which relies on the interviewer's clinical judgment, as the criterion. A roadblock to using clinician ratings as outcome criteria is that many patients had left therapy at the 5-year follow-up, making collection of accurate clinician ratings impossible. As such, future work might broaden criteria to include method-neutral outcomes such as hospitalizations or suicide attempts.

Relatedly, it should also be noted that the use of *any* prospective criteria presumes at least some diagnostic stability. After all, if the diagnoses and associated impairments were transitory, one would expect no relationship between a diagnosis and functioning over any time interval. The current study helps to demonstrate that PD diagnoses, provided by any source or method, do predict subsequent impairment. This corroborates the notion that PD diagnoses have some temporal stability but that their associated impairments may be even more durable (Zanarini, Frankenburg, Reich, & Fitzmaurice, 2010). The degree to which participants' functioning improved through treatment would diminish the predictive utility of a baseline diagnosis. Importantly, though, this would not favor or disfavor any source or method more than another and thus should not affect our findings.

As in any longitudinal study, some participants dropped out. We chose the 5-year interval to balance available sample size with meaningful duration, yet ideally we would like to have had functioning data on all participants. Concern about this potential limitation is tempered by independent samples t tests revealing no significant differences between attriters and those retained on baseline functioning or any other measure.

Finally, although our sample included representation from minority groups, the overall ethnic composition was predominantly White, potentially limiting the generalizability of these data. Future research that examines potential differences with regard to ethnicity and other demographic variables would be helpful.

Conclusions

Our findings contribute further evidence that PD diagnoses made by treating clinicians agree poorly with semistructured in-

terviews and self-report questionnaires. Most importantly, our novel findings provide evidence that the latter two methods have greater utility than clinicians' PD diagnoses for predicting psychosocial functioning over 5-year prospective follow-up. These findings underscore the advantages of incorporating established semistructured interviews and self-report questionnaires into routine clinical diagnostic practices.

References

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

Blashfield, R. K., & Herkov, M. J. (1996). Investigating clinician adherence to diagnosis by criteria: A replication of Morey and Ochoa (1989). *Journal of Personality Disorders, 10,* 219–228. doi:10.1521/pedi.1996.10.3.219

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6,* 284–290. doi:10.1037/1040-3590.6.4.284

Clark, L. A., Simms, L. J., Wu, K. D., & Casillas, A. (in press). *Manual for the Schedule for Nonadaptive and Adaptive Personality (SNAP-2)*. Minneapolis: University of Minnesota Press.

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,* 208–218. doi:10.1521/pedi.17.3.208.22148

Dreessen, L., & Arntz, A. (1999). Personality disorders have no excessively negative impact on therapist-rated therapy process in the cognitive and behavioural treatment of Axis I anxiety disorders. *Clinical Psychology & Psychotherapy, 6,* 384–394. doi:10.1002/(SICI)1099-0879(199911)6:5<384::AID-CPP218>3.0.CO;2-8

Elkin, I., Parloff, M. B., Hadley, S. W., & Autry, J. H. (1985). NIMH Treatment of Depression Collaborative Research Program: Background and research plan. *Archives of General Psychiatry, 42,* 305–316. doi:10.1001/archpsyc.1985.01790260103013

Fridell, M., & Hesse, M. (2006). Clinical diagnosis and SCID-II assessment of DSM-III-R personality disorders. *European Journal of Psychological Assessment, 22,* 104–108. doi:10.1027/1015-5759.22.2.104

Ganellen, R. J. (2007). Assessing normal and abnormality personality functioning: Strengths and weaknesses of self-report, observer, and performance-based methods. *Journal of Personality Assessment, 89,* 30–40. doi:10.1080/00223890701356987

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12,* 19–30. doi:10.1037/1040-3590.12.1.19

Gunderson, J. G., Shea, M. T., Skodol, A. E., McGlashan, T. H., Morey, L. C., Stout, R. L., . . . Keller, M. B. (2000). The collaborative longitudinal personality disorders study: Development, aims, design, and sample characteristics. *Journal of Personality Disorders, 14,* 300–315. doi:10.1521/pedi.2000.14.4.300

Hopwood, C. J., Morey, L. C., Edelen, M. O., Shea, M. T., Grilo, C. M., Sanislow, C. A., . . . Skodol, A. E. (2008). A comparison of interview and self-report methods for the assessment of borderline personality disorder criteria. *Psychological Assessment, 20,* 81–85. doi:10.1037/1040-3590.20.1.81

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,* 557–570. 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,* 170–178. doi:10.1016/0010-440X(89)90070-9

Keller, M. B., Lavori, P. W., Friedman, B., Nielsen, E., Endicott, J., Mcdonaldscott, P., & Andreasen, N. C. (1987). The Longitudinal Interval Follow-up Evaluation: A comprehensive method for assessing outcome in prospective longitudinal studies. *Archives of General Psychiatry, 44,* 540–548. doi:10.1001/archpsyc.1987.01800180050009

Klein, D. N. (2003). Patients' versus informants' reports of personality disorders in predicting 7 1/2-year outcome in outpatients with depressive disorders. *Psychological Assessment, 15,* 216–222. doi:10.1037/1040-3590.15.2.216

Krueger, R. F., Derringer, J., Markon, K. E., Watson, D., & Skodol, A. E. (2012). Initial construction of a maladaptive personality trait model and inventory for DSM-5. *Psychological Medicine, 42,* 1879–1890. doi:10.1017/S0033291711002674

Loranger, A. W., Lenzenweger, M. F., Gartner, A. F., Susman, V. L., Herzig, J., Zammit, G. K., . . . Young, R. C. (1991). Trait-state artifacts and the diagnosis of personality disorders. *Archives of General Psychiatry, 48,* 720–728. doi:10.1001/archpsyc.1991.01810320044007

McDermut, W., & Zimmerman, M. (2005). Assessment instruments and standardized evaluation. In J. Oldham, A. E. Skodol, & D. Bender (Eds.), *The American Psychiatric Publishing textbook of personality disorder* (pp. 89–102). Washington, DC: American Psychiatric Publishing.

Meehl, P. E. (1954). *Clinical versus statistical prediction; a theoretical analysis and a review of the evidence*. Minneapolis: University of Minnesota Press. doi:10.1037/11281-000

Morey, L. C., Blashfield, R. K., Webb, W. W., & Jewell, J. (1988). MMPI scales for DSM-III personality disorders: A preliminary validation study. *Journal of Clinical Psychology, 44,* 47–50. doi:10.1002/1097-4679(198801)44:1<47::AID-JCLP2270440110>3.0.CO;2-R

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,* 180–192. doi:10.1521/pedi.1989.3.3.180

Morey, L. C., Shea, M. T., Markowitz, J. C., Stout, R. L., Hopwood, C. J., Gunderson, J. G., . . . Skodol, A. E. (2010). State effects of major depression on the assessment of personality and personality disorder. *American Journal of Psychiatry, 167,* 528–535. doi:10.1176/appi.ajp.2009.09071023

Nestadt, G., Di, C., Samuels, J. F., Cheng, Y.-J., Bienvenu, O. J., Reti, I. M., . . . Bandeen-Roche, K. (2012). Concordance between personality disorder assessment methods. *Psychological Medicine, 42,* 657–667. doi:10.1017/S0033291711001632

Oltmanns, T. F., & Turkheimer, E. (2009). Person perception and personality pathology. *Current Directions in Psychological Science, 18,* 32–36. doi:10.1111/j.1467-8721.2009.01601.x

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

Pilkonis, P. A., Hallquist, M. N., Morse, J. Q., & Stepp, S. D. (2011). Striking the (im)proper balance between scientific advances and clinical utility: Commentary on the DSM-5 proposal for personality disorders. *Personality Disorders: Theory, Research, and Treatment, 2,* 68–82. doi:10.1037/a0022226

Pilkonis, P. A., Heape, C. L., Ruddy, J., & Serrao, P. (1991). Validity in the diagnosis of personality disorders: The use of the LEAD standard. *Psychological Assessment, 3,* 46–54. doi:10.1037/1040-3590.3.1.46

Rettew, D. C., Lynch, A. D., Achenbach, T. M., Dumenci, L., & Ivanova, M. Y. (2009). Meta-analyses of agreement between diagnoses made from clinical evaluations and standardized diagnostic interviews. *International Journal of Methods in Psychiatric Research, 18,* 169–184. doi:10.1002/mpr.289

Rossi, G., Van den Brande, I., Tobac, A., Sloore, H., & Hauben, C. (2003). Convergent validity of the MCMI-III personality disorder scales and the MMPI-2 scales. *Journal of Personality Disorders, 17,* 330–340. doi:10.1521/pedi.17.4.330.23970

Rottman, B. M., Ahn, W.-k., Sanislow, C. A., & Kim, N. S. (2009). Can clinicians recognize DSM-IV personality disorders from five-factor model descriptions of patient cases? *American Journal of Psychiatry, 166,* 427–433. doi:10.1176/appi.ajp.2008.08070972

Samuel, D. B., Hopwood, C. J., Ansell, E. B., Morey, L. C., Sanislow, C., Markowitz, J. C., . . . Grilo, C. M. (2011). Comparing the temporal stability of self-report and interview assessed personality disorder. *Journal of Abnormal Psychology, 120,* 670–680. doi:10.1037/a0022647

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,* 721–745. doi:10.1521/pedi.2010.24.6.721

Shea, M. T., Glass, D. R., Pilkonis, P. A., Watkins, J. T., & Docherty, J. P. (1987). Frequency and implications of personality disorders in a sample of depressed outpatients. *Journal of Personality Disorders, 1,* 27–42. doi:10.1521/pedi.1987.1.1.27

Shea, M. T., Pilkonis, P. A., Beckham, E., Collins, J. F., Elkin, I., Sotsky, S. M., & Docherty, J. P. (1990). Personality disorders and treatment outcome in the NIMH treatment of depression collaborative research program. *American Journal of Psychiatry, 147,* 711–718.

Shedler, J., Beck, A., Fonagy, P., Gabbard, G. O., Gunderson, J., Kernberg, O., . . . Westen, D. (2010). Personality disorders in DSM-5. *American Journal of Psychiatry, 167,* 1026–1028. doi:10.1176/appi.ajp.2010.10050746

Shedler, J., & Westen, D. (2004). Refining personality disorder diagnosis: Integrating science and practice. *American Journal of Psychiatry, 161,* 1350–1365. doi:10.1176/appi.ajp.161.8.1350

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 [Review]. *Journal of Personality Disorders, 25,* 136–169. doi:10.1521/pedi.2011.25.2.136

Skodol, A. E., Clark, L. A., Bender, D. S., Krueger, R. F., Morey, L. C., Verheul, R., . . . Oldham, J. M. (2011). Proposed changes in personality and personality disorder assessment and diagnosis for DSM-5 Part I: Description and rationale. *Personality Disorders: Theory, Research, and Treatment, 2,* 4–22. doi:10.1037/a0021891

Weissman, M. M., & Bothwell, S. (1976). Assessment of social adjustment by patient self-report. *Archives of General Psychiatry, 33,* 1111–1115. doi:10.1001/archpsyc.1976.01770090101010

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. (2004). *The Clinical Diagnostic Interview*. Atlanta, GA: Emory University, Departments of Psychology and Psychiatry and Behaviorial Sciences.

Westen, D., DeFife, J. A., Bradley, B., & Hilsenroth, M. J. (2010). Prototype personality diagnosis in clinical practice: A viable alternative for DSM–5 and ICD–11. *Professional Psychology: Research and Practice, 41,* 482–487. doi:10.1037/a0021555

Westen, D., & Muderrisoglu, S. (2003). Assessing personality disorders using a systematic clinical interview: Evaluation of an alternate to structured interviews. *Journal of Personality Disorders, 17,* 351–369. doi:10.1521/pedi.17.4.351.23967

Westen, D., & Muderrisoglu, S. (2006). Clinical assessment of pathological personality traits. *American Journal of Psychiatry, 163,* 1285–1287. doi:10.1176/appi.ajp.163.7.1285

Westen, D., & Shedler, J. (1999). Revising and assessing Axis II, Part I: Developing a clinically and empirically valid assessment method. *American Journal of Psychiatry, 156,* 258–272. doi:10.1521/pedi.2000.14.4.291

Westen, D., Shedler, J., & Bradley, R. (2006). A prototype approach to personality disorder diagnosis. *American Journal of Psychiatry, 163,* 846–856. doi:10.1176/appi.ajp.163.5.846

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. doi: 10.1176/appi.ajp.2011.11020274

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

Widiger, T. A. (2011). A shaky future for personality disorders. *Personality Disorders: Theory, Research, and Treatment, 2,* 54–67. doi:10.1037/a0021855

Widiger, T. A., & Boyd, S. (2009). Assessing personality disorders. In J. Butcher (Ed.), *Oxford handbook of personality assessment* (pp. 336–363). New York, NY: Oxford University Press. doi:10.1093/oxfordhb/9780195366877.013.0018

Zanarini, M. C., Frankenburg, F. R., Reich, D. B., & Fitzmaurice, G. (2010). The 10-year course of psychosocial functioning among patients with borderline personality disorder and Axis II comparison subjects. *Acta Psychiatrica Scandinavica, 122,* 103–109. doi:10.1111/j.1600-0447.2010.01543.x

Zanarini, M. C., Frankenburg, F. R., Sickel, A. E., & Yong, L. (1996). *The Diagnostic Interview for DSM-IV Personality Disorders (DIPD-IV)*. Belmont, MA: McLean Hospital.

Zanarini, M. C., Skodol, A. E., Bender, D., Dolan, R., Sanislow, C., Schaefer, E., . . . Gunderson, J. G. (2000). The collaborative longitudinal personality disorders study: Reliability of Axis I and II diagnoses. *Journal of Personality Disorders, 14,* 291–299. doi:10.1521/pedi.2000.14.4.291

Zimmerman, M. (1994). Diagnosing personality disorders: A review of issues and research methods. *Archives of General Psychiatry, 51,* 225–245. doi:10.1001/archpsyc.1994.03950030061006

Zimmerman, M. (2011). A critique of the proposed prototype rating system for personality disorders in DSM-5. *Journal of Personality Disorders, 25,* 206–221. doi:10.1521/pedi.2011.25.2.206

Zimmerman, M., & Mattia, J. I. (1999). Differences between clinical and research practices in diagnosing borderline personality disorder. *American Journal of Psychiatry, 156,* 1570–1574.

Received December 1, 2011
Revision received December 10, 2012
Accepted March 15, 2013 ■