Journal of Abnormal Psychology
2011, Vol. 120, No. 3, 670–680

0021-843X/11/$12.00 DOI: 10.1037/a0022647

Comparing the Temporal Stability of Self-Report and Interview Assessed Personality Disorder

Douglas B. Samuel
Yale University School of Medicine and VA New England Mental Illness Research, Education and Clinical Center

Christopher J. Hopwood
Michigan State University

Emily B. Ansell
Yale University School of Medicine

Leslie C. Morey
Texas A&M University

Charles A. Sanislow
Wesleyan University

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

Shirley Yen
Brown University Alpert Medical School

M. Tracie Shea
Department of Veterans Affairs and Brown University Alpert Medical School

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

Carlos M. Grilo
Yale University School of Medicine

Findings from several large-scale, longitudinal studies over the last decade have challenged the long-held assumption that personality disorders (PDs) are stable and enduring. However, the findings, including those from the Collaborative Longitudinal Personality Disorders Study (CLPS; Gunderson et al., 2000), rely primarily on results from semistructured interviews. As a result, less is known about the stability of PD scores from self-report questionnaires, which differ from interviews in important ways (e.g., source of the ratings, item development, and instrument length) that might increase temporal stability. The current study directly compared the stability of the *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.; *DSM–IV*) PD constructs assessed via the Schedule for Nonadaptive and Adaptive Personality (SNAP–2; Clark, Simms, Wu, & Casillas, in press) with those from the Diagnostic Interview for *DSM–IV* Personality Disorders (Zanarini, Frankenburg, Sickel, & Yong, 1996) over 2 years in a sample of 529 CLPS participants. Specifically, we compared dimensional and categorical representations from both measures in terms of rank-order and mean-level stability. Results indicated that the dimensional scores from the self-report questionnaire had significantly greater rank-order (mean $r = .69$ vs. .59) and mean-level (mean $d = 0.21$ vs. 0.30) stability. In contrast, categorical diagnoses from the two measures evinced comparable rank-order (mean $\kappa = .38$ vs. .37) and mean-level stability (median prevalence rate decrease of 3.5% vs. 5.6%). These findings suggest the stability of PD constructs depends at least partially on the method of assessment and are discussed in the context of previous research and future conceptualizations of personality pathology.

Keywords: stability, personality disorder, retest, consistency, self-report

This article was published Online First March 28, 2011.

Douglas B. Samuel, Department of Psychiatry, Yale University School of Medicine and VA New England Mental Illness Research, Education and Clinical Center; Christopher J. Hopwood, Department of Psychology, Michigan State University; Emily B. Ansell and Carlos M. Grilo, Department of Psychiatry, Yale University School of Medicine; Leslie C. Morey, Department of Psychology, Texas A&M University; Charles A. Sanislow, Department of Psychology, Wesleyan University; John C. Markowitz, Department of Psychiatry, New York State Psychiatric Institute and Columbia University College of Physicians & Surgeons; Shirley Yen, Department of Psychiatry and Human Behavior, Brown University Alpert Medical School; M. Tracie Shea, Department of Veterans Affairs and Department of Psychiatry and Human Behavior, Brown University Alpert Medical School; Andrew E. Skodol, The Sunbelt Collaborative, Tucson, Arizona, and Department of Psychiatry, University of Arizona College 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 (awarded to Charles A. Sanislow). This article has been reviewed and approved by the Publications Committee of the Collaborative Longitudinal Personality Disorders Study. No conflicts of interest are represented by any of the authors. We thank Karl L. Wuensch and James H. Steiger for their input on statistical comparisons.

Correspondence concerning this article should be addressed to Douglas B. Samuel, who is now at Department of Psychological Sciences, Purdue University, 703 Third Street, West Lafayette, IN 47097-2081. E-mail: dbsamuel@purdue.edu

The American Psychiatric Association's (APA's; 2000) *Diagnostic and Statistical Manual of Mental Disorders* (text revision; *DSM–IV–TR*) defined a personality disorder (PD) as "an enduring pattern of inner experience and behavior that . . . is stable over time" (p. 685). Nonetheless, over the past decade, findings from longitudinal studies have cast doubt over whether PD constructs are, in fact, defined by temporal stability (Cohen, Crawford, Johnson, & Kasen, 2005; Skodol et al., 2005; Zanarini, Frankenburg, Hennen, Reich, & Silk, 2005). Results from these studies have suggested that prevalence of PDs in community participants decreased steadily from adolescence to early adulthood (Johnson et al., 2000).

Another potentially surprising finding from two studies in clinical samples was a relatively high rate of remission. Zanarini et al. (2005) indicated that nearly three quarters of the individuals with diagnoses of borderline personality disorder (BPD) no longer met criteria by 6-year follow-up and concluded that "BPD is relatively stable over time compared to mood disorders" but, in contrast to the *DSM* definition, is "mutable over more sustained periods of time" (p. 513). Similarly, the Collaborative Longitudinal Personality Disorders Study (CLPS) found that less than half of PD patients remained at diagnostic threshold over periods of 1 year (Shea et al., 2002) and that most experienced remission (defined as 12 consecutive months with no more than two diagnostic criteria) within the first 2 years of the study (Grilo et al., 2004).

The above findings primarily concern absolute changes and certainly suggest notable mean-level decreases in PD scores both within and across individuals. However, one can also examine the relative, or rank-order, stability of these scores across individuals (e.g., the correlations between scores at different time points; Roberts, Wood, & Caspi, 2008). Previous results from CLPS have suggested that dimensional scores demonstrate greater rank-order stability than do categorical diagnoses (Grilo et al., 2004; Morey et al., 2007). It is worth noting, though, that Morey et al. (2007) also indicated that even the rank-order stability of dimensional PD scores was significantly lower than that for scores of trait models of general personality functioning. However, this particular comparison is confounded because the PD constructs were assessed via semistructured interview, whereas the trait model scores come from a self-report questionnaire.

In fact, one commonality among these large-scale longitudinal studies is their focus on PD stability as assessed by semistructured interviews. This is regarded as a methodological strength (McDermut & Zimmerman, 2005; Widiger & Samuel, 2005) because interviews rely on trained clinical assessors who carefully document the presence or absence of each diagnostic criterion through a series of open-ended questions, often taking into account participant behaviors during the interview (Rogers, 2001). Nonetheless, as a result of this practice, less is known about the temporal stability of PDs assessed via self-report questionnaires, which are commonly used in research and clinical settings (Lenzenweger, Loranger, Korfine, & Neff, 1997; Widiger & Samuel, 2005). Although there are potential disadvantages to studying the temporal stability of PDs via a self-report questionnaire, such as susceptibility to bias from Axis I symptoms (e.g., Piersma, 1989; Zimmerman, 1994; but see also Morey et al., 2010), there are also compelling reasons why it is informative and useful.

It is important to examine the stability of PD scores on self-report questionnaires because they provide information that is usefully different from other assessment methods. For example, Hopwood et al. (2008) demonstrated that a self-report questionnaire and an interview measure of borderline PD each incremented the other in predicting functional impairment. More specifically, each method had unique strengths: Self-report questionnaires fared better for diagnostic criteria that were experiential in nature (e.g., identity disturbance), whereas the interview measure was superior for more externally observable indicators (e.g., impulsivity).

A fundamental difference between self-report questionnaires and semistructured interviews that might influence their temporal consistency is that interviews compound sources of potential score variance. The same individual completes self-report measures at each time point; hence, the only source of score variability is a difference in that individual's reporting. In contrast, interviews require the judgment of a second person and thus inherently contain not only reporting variability (e.g., the interviewee answers the same question differently) but also variable perception (e.g., the interviewee gives the same answer, yet the interviewer scores it differently) over multiple assessments. Compounding this even further, different clinicians often administer semistructured interviews at subsequent assessments (Zimmerman, 1994). Previous findings have demonstrated that scoring variability across interviewers influences tests of cognitive ability (e.g., Hopwood & Richards, 2005), and this might also be true for diagnostic interviews for PDs. Thus, temporal consistency might be higher for a self-report questionnaire simply because error variance attenuates stability for the semistructured interview.

Another important contribution of self-report inventories is that their development differs somewhat from the development of interview measures. Self-report questionnaires typically are derived through an iterative process (i.e., Clark & Watson, 1995) that begins with writing many candidate items and administering them to large samples of participants. Although interview measures also undergo rigorous development, such detailed testing is more difficult because of the time cost of administering interview items. This same property results in many self-report questionnaires having more items assessing each PD than do semistructured interviews, which typically include one scored item per diagnostic criterion but allow follow-up questions at the interviewer's discretion. The greater length of self-report questionnaires might allow them to obtain a more fine-grained assessment of each PD construct than semistructured interviews would. Relatively brief instruments, which include only a few items, might be perfectly acceptable for assessing narrow constructs. However, as the breadth of the construct increases, a greater number of items might be required to capture it adequately. This seems particularly relevant to PDs, which are considered quite heterogeneous in nature (Trull & Durrett, 2005). Thus, having multiple scored items to assess a diagnostic criterion might provide an advantage for the self-report questionnaire. At the very least, having more scored items yields a greater range of possible dimensional scores, which might increase temporal stability in and of itself. In any event, a direct comparison of the relative temporal consistency of self-report questionnaires and semistructured interviews would help to determine whether existing findings reflect the stability of the PD constructs themselves or properties of the method of assessment.

One longitudinal study (the Longitudinal Study of Personality Disorders; LSPD; Lenzenweger, 2006) has provided stability results from a self-report questionnaire as well as a semistructured

interview. Among a sample of college undergraduates, Lenzenweger and his colleagues identified 134 students who met criteria for at least one *DSM–III–R* (APA, 1987) PD and another 124 with virtually no PD pathology (i.e., fewer than 10 criteria across the PDs). These individuals were reassessed twice using a semistructured interview (the International Personality Disorder Examination; IPDE; Loranger, 1999) and a self-report PD measure (the Millon Clinical Multiaxial Inventory—II; MCMI–II; Millon, 1987) over the next 3 years (Lenzenweger, 1999). When the PD scores from the IDPE were considered dimensionally, the rank-order stability coefficients ranged from .44 (avoidant) to .74 (schizoid), with a mean of .57. The self-reported scores from the MCMI–II obtained coefficients ranging from a low of .63 to a high of .76, with a mean of .71 (Lenzenweger, 1999). Although these values were not tested against each other, Lenzenweger (1999) noted that stability was higher for self-report questionnaires than for semistructured interviews. Lenzenweger also noted significant, albeit relatively small, mean level decreases for many of the PD scores on both instruments over time, with most of this change occurring in the first reassessment. Appreciable differences regarding the mean-level stability of the dimensional scores were not noted between the two assessment methods. Finally, indices of categorical agreement (e.g., kappa) for individual PDs could not be calculated because of the low base rates within the sample.

Although informative, these LSPD results are potentially limited because the individuals assessed, although endorsing significant ranges of PD symptoms, were nonclinical university students. There are conceptual advantages to studying PDs within the general population. Community samples might provide a more naturalistic picture of the pathology as it exists in nature compared with a sample influenced by whether a given individual decides to seek treatment. Nonetheless, the use of clinical samples also has appreciable advantages: It facilitates obtaining the complete range of possible pathology as well as higher rates of individual diagnoses. This is particularly useful statistically and conceptually as it oversamples the upper ranges of each PD construct, which is, by definition, the portion of greatest clinical interest.

Trull and Goodwin (1993) examined the temporal stability of *DSM–III–R* PD scores within a clinical sample and reported the mean rank-order stability coefficient across the 10 PDs was .61 for an interview measure, whereas two self-report questionnaires obtained values of .75 and .65. These values appear similar to those reported by Lenzenweger (1999), bolstering support for the notion that PD scores from self-report questionnaires might show higher rank-order stability than those from semistructured interviews. Nonetheless, the results of Trull and Goodwin reflect a sample of only 44 psychiatric outpatients who were reassessed over a 6-month interval. It would be useful to replicate and extend these results using a larger clinical sample over a longer duration. In addition, the findings from both Trull and Goodwin (1993) and Lenzenweger (1999) concern assessments of *DSM–III–R* PD constructs. It would be useful to update these findings for *DSM–IV* constructs.

Surprisingly few studies have even examined the temporal stability of self-report questionnaires assessing the *DSM–IV* PDs. At least 10 self-report measures provide an assessment of the PD constructs (Widiger & Boyd, 2009) and most have been used to examine temporal stability in at least one study. However, most studies used older versions of these instruments, assessing PDs from prior editions of the diagnostic manual. In fact, a literature search revealed only six studies that have examined the stability of a self-report questionnaire assessing the *DSM–IV* PD constructs. Even this literature might be deemed limited, as studies typically considered only one type of stability (i.e., rank order) and most used nonclinical samples (e.g., Okada & Oltmanns, 2009) or reassessed over very brief intervals (e.g., 1–6 weeks; Millon, 1994; Ottosson, Grann, & Kullgren, 2000; Piersma & Boes, 1997).

Perhaps the most relevant data were provided by Craig and Olson (1998), who administered the MCMI–III to 35 African American men in an inpatient substance use treatment facility and reported test–retest correlations for the 10 *DSM–IV* PDs over a 6-month interval. Craig and Olson reported rank-order stability coefficients ranging from .52 (schizotypal) to .83 (dependent), with a median of .69. Although these studies provide information regarding the stability of the self-reported *DSM–IV* PDs, they are limited because they examined stability in relatively small samples of individuals engaged in active treatment for Axis I disorders. It would be useful to examine stability in a sample with a greater range of personality pathology. Finally, the use of the MCMI–III also could be considered problematic for studying stability as there is extensive item overlap between the scales, such that changes on a single item would alter the stability of more than one PD.

The participants in CLPS completed a self-report questionnaire assessing *DSM–IV* PDs: the Schedule for Nonadaptive and Adaptive Personality—2 (SNAP–2; Clark, Simms, Wu, & Casillas, in press). The SNAP–2 was derived through analyses of maladaptive personality symptoms (see Clark, 1993) and assesses three broad temperaments (e.g., disinhibition vs. constraint) and 12 traits that fall beneath these domains (e.g., impulsivity, propriety, and workaholism). An emerging literature supports the reliability and validity of the SNAP–2 temperament and trait scales (e.g., Simms & Clark, 2006), which have correlated well with other measures of personality pathology (e.g., Reynolds & Clark, 2001) and have shown predictable relationships with PD pathology (Morey et al., 2003).

The SNAP–2 also includes scales assessing the 10 *DSM–IV–TR* PDs. These PD scales range in length from 19 (avoidant) to 34 (antisocial) items, with a median of 25. Each diagnostic criterion is assessed by at least two items, which allows the PD scales to be scored categorically (i.e., meeting a sufficient number of criteria) or dimensionally. Although items overlap between the PD scales and the trait and temperament scales, novel items were developed for the PD scales when extant items did not assess specific criteria well. Thus, the PD scales were constructed for a different purpose and contain unique items not scored on any trait or temperament scale. In addition, all of the SNAP–2 PD scales are nonoverlapping, as items are scored for only a single PD. Previous CLPS studies have examined the temporal stability of the temperament and trait scales (Morey et al., 2007) but not the SNAP–2 PD scales.

In fact, only a single study has described the temporal stability of SNAP PD scores (Melley, Oltmanns, & Turkheimer, 2002). Melley et al. (2002) reported that test–retest correlations of dimensional PD scores over a 9-month interval ranged from .59 (schizotypal) to .84 (antisocial), with a median of .75. However, this study was conducted within a sample of undergraduates and used the original version of the SNAP (Clark, 1993), which assesses the *DSM–III–R* PDs. The temporal stability of the SNAP–2 PD scales has yet to be examined.

Beyond general qualities of self-report instruments, there are two particular advantages of studying stability of self-reported PD scores in the CLPS sample. First, the scores on the semistructured interview (i.e., Diagnostic Interview for *DSM–IV* Personality Disorders; DIPD-IV; Zanarini, Frankenburg, Sickel, & Yong, 1996) were used to determine inclusion in the CLPS sample. This strategy ensured an adequate representation of the PDs but, by definition, also slanted the sample toward individuals with extreme DIPD-IV scores and potentially increased false positives (Chmielewski & Watson, 2009). In fact, a majority of the change (i.e., decrease) observed for the DIPD-IV scores occurred during the study's first six months (Grilo et al., 2004), which some have interpreted as regression to the mean (Clark, 2005). The SNAP–2 PD scales were administered at CLPS baseline assessment but infrequently used for inclusion decisions (Morey et al., 2003) and hence should be less prone to these issues. Another strength of the CLPS sample for testing temporal stability is that participants, although mostly treatment seeking at study onset, did not necessarily receive treatment throughout follow-up. This allows a more naturalistic look at the stability of the PD constructs that is at least partially independent of the effects of active treatment.

The current study builds on previous research in several important ways. First, it investigates the temporal stability of the *DSM–IV* PDs assessed via self-report questionnaire in a large, clinical sample with appreciable rates of PD diagnoses. In addition, the current study extends the previous literature on the stability of self-reported PD pathology by examining rank-order and mean-level stabilities of both dimensional and categorical representations of the PD constructs. Finally, it explicitly compares, using these metrics, the relative stability of PD ratings from a self-report questionnaire with those from a semistructured interview within the same sample.

Method

Participants

Participants for this study were drawn from the 668 recruited from multiple clinical sites for the CLPS. Participants underwent clinical diagnostic interviews and completed self-report instruments as part of a standardized assessment process (Gunderson et al., 2000). They were assigned to one of four PD groups (borderline, avoidant, schizotypal, and obsessive-compulsive) or to the group of participants with major depressive disorder but no PD diagnosis on the basis of reliably administered diagnostic interviews. Additional details regarding recruitment, screening, and diagnostic procedures have been previously published (Gunderson et al., 2000). Participants were not excluded on the basis of the presence of other nonstudy PDs, and they received an average of 2.1 PD diagnoses (McGlashan et al., 2000). To limit the effect of sampling on our results, we followed the same procedures as Morey et al. (2003) and confined our sample to participants for whom the SNAP–2 was not used to establish diagnostic assignment. This subsample contains 432 participants assigned to one of four primary PD groups. The number of participants in each PD group was 40 for schizotypal, 139 for borderline, 128 for avoidant, and 125 for obsessive-compulsive. There was also a comparison group of 97 individuals who met criteria for major depressive disorder but had no PD diagnosis, bringing the total sample to 529. Participants were not included in the major depressive disorder group if they had 15 or more PD symptoms or came within two criteria of any PD diagnosis. This is important for the current analyses, as it increases the variability in the SNAP–2 PD scale scores. The sample used in this study was primarily Caucasian (76%) and female (64%), with an average age at intake of 32.7 years (SD = 8.1).

The SNAP–2 was administered at 6 months, 1 year, and 2 years after baseline and then biennially throughout CLPS. The DIPD-IV was administered at baseline and biannually. Because of attrition or failure to complete the SNAP–2, the total sample of 529 decreased to 356 by the 2-year assessment. Although additional data are available concerning the stability of these constructs through 10 years, attrition further limited the available sample at these latter points. Thus, we considered only data through 2 years to maximize the available sample size and provide the most robust estimates of stability (Watson, 2004).

Instruments

SNAP–2 (Clark, Simms, Wu, & Casillas, in press). Comprising 390 true–false statements, the SNAP–2 provides a self-report assessment of a dimensional model of personality pathology and the *DSM–IV–TR* PDs (APA, 2000). The latter scales dimensionally assess the PDs and range in length from 19 (avoidant) to 34 (antisocial) items. In the current sample, the SNAP–2 PD scale internal consistencies ranged from .69 (obsessive-compulsive) to .88 (avoidant), with an overall median of .83. The SNAP–2 PD scores correlate strongly with those from other self-report PD inventories (see Widiger & Boyd, 2009) and scores from a structured PD interview (Samuel et al., 2010).

DIPD-IV (Zanarini et al., 1996). The DIPD-IV is a semistructured diagnostic interview for assessing PD. Each of the criteria for all PD diagnoses is assessed with one or more questions, which are then 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 present and pervasive for at least two years and be characteristic of the person for most of his or her adult life to be counted toward a diagnosis. In the present study, interrater reliability (based on 84 pairs of raters) kappa coefficients for PD ranged from .58 to 1.00 (Zanarini et al., 2000).

Structured Clinical Interview for *DSM–IV* Axis I Disorders—Patient Version (SCID–I/P; First, Spitzer, Gibbon, & Williams, 1996). The SCID–I/P is a semistructured diagnostic interview for assessing current and lifetime Axis I psychiatric disorders. In the present study, kappa coefficients for interrater reliability for Axis I diagnoses ranged from .57 to 1.0; kappa for major depressive disorder was .80 (Zanarini et al., 2000).

Data Analyses

We calculated the temporal stability of the PD scores in terms of both rank-order stability and mean-level change. In addition, because we were interested in both the dimensional and the categorical representations of the PD constructs, we computed these values separately. Finally, to facilitate the comparison across methods, we computed these sets of values for the self-report scores from SNAP–2 and the interview scores from the DIPD-IV.

This creates four separate points of comparison that fit a 2 × 2 matrix, with dimensional and categorical scoring across the rows and rank-order stability and mean-level change down the columns.

For the first cell, featuring the rank-order stability of the dimensional scores, we computed Pearson correlations between scores at baseline and the 2-year retest. These correlations indicate the degree to which the rank ordering of participants remained constant. We then compared the resulting values for each PD across the methods using Steiger's (1980) method for comparing dependent correlations. This method produces a *z* score value that determines whether the differences between the correlations are significant.

We also examined the mean-level change to assess how much dimensional PD scores changed, on average, over time. We used the means and standard deviations at baseline and 2 years to compute effect size estimates (Cohen's *d*). These effect size estimates were standardized to the baseline assessment for the 356 participants with all data available and represent the magnitude of change from baseline. To index whether the mean-level change was significantly different between the self-report and interview methods, we then computed difference scores between the two assessments (e.g., SNAP–2 avoidant PD score at baseline subtracted from SNAP–2 avoidant PD score at Year 2). Because the two measures have different numbers of items, we first equated them so that these difference scores shared the same metric. We then compared the change scores for each instrument using a paired-samples *t* test.

We next investigated the rank-order stability of the categorical representations of the PDs for each instrument using kappa coefficients. These values indicate the diagnostic agreement between diagnoses assigned at baseline and follow-up within each instrument. The kappa coefficients for each PD were compared between the two methods using a bootstrapping procedure (with 1,000 samples) to produce a 95% confidence interval around the kappa values for the SNAP–2 and DIPD-IV (Vanbelle & Albert, 2008). The presence of nonoverlapping confidence intervals is more conservative than null hypothesis testing but is the only method by which to compare these same-sample kappas.

Finally, we also sought to examine the mean-level change of the categorical diagnoses. Because these rely on the same cross-tabulations on which the kappa coefficients were based, we computed the percentage of individuals who met criteria according to each instrument at each time point. This provides the most equivalent method of determining whether the sample, on average, demonstrated change in terms of the categorical diagnoses. We are unaware of any method that permits null hypothesis statistical testing for these percentages.

Results

Stability of Dimensional PD Scores

Table 1 provides both the rank-order and the mean-level stability of the dimensional scores on the SNAP–2 and DIPD-IV. The second column of Table 1 presents the Pearson correlations between the baseline and 2-year assessments for the SNAP–2 PD scales. The third column presents the same values calculated using criterion counts from the DIPD-IV. Beneath each column are median and mean values, with the latter calculated using Fisher's *r*-to-*z* transformation, averaging, then converting back to correlations. Statistical comparisons conducted using Steiger's (1980) method indicated that the dimensional scores from the SNAP–2 had significantly higher rank-order stability than did those from the DIPD-IV for six PDs. There were no PDs for which the DIPD-IV scores were more stable.

Table 1 also presents the mean-level change on the dimensional PD scores. Paired-samples *t* tests indicate that all PDs showed a significant decrease. To ease comparison between methods, we present these values for the dimensional scores in terms of Cohen's *d*, such that each column indicates the effect size change from baseline to the 2-year assessment. For instance, the SNAP–2 borderline PD scores decreased by an effect size of 0.31, whereas the DIPD-IV borderline scores decreased by 0.43. Mean and median effect sizes across the 10 PDs appear below the columns. Paired sample *t* tests of the difference scores indicated that the DIPD scores decreased significantly more than SNAP–2 scores for

Table 1
Stabilities for Dimensional Representations of the Personality Disorders

	Rank order (*r*)			Mean level (*d*)		
Personality disorder	SNAP–2	DIPD–IV	*z*	SNAP–2	DIPD–IV	*t*
Paranoid	0.74	0.57	4.54***	−0.23	−0.21	−0.96
Schizoid	0.68	0.44	4.83***	−0.15	−0.20	−0.42
Schizotypal	0.71	0.70	0.22	−0.31	−0.27	−1.05
Antisocial	0.84	0.84	0.15	−0.06	−0.09	0.35
Borderline	0.67	0.63	1.12	−0.31	−0.43	4.24***
Histrionic	0.70	0.45	5.07***	−0.13	−0.35	2.56*
Narcissistic	0.63	0.49	2.77**	−0.11	−0.27	1.99*
Avoidant	0.68	0.65	0.85	−0.24	−0.37	3.67***
Dependent	0.61	0.45	3.06***	−0.26	−0.34	0.57
Obsessive-compulsive	0.61	0.50	2.01*	−0.29	−0.46	5.01***
Mdn	0.68	0.54		−0.24	−0.31	
M[a]	0.69	0.59		−0.21	−0.30	

Note. Values presented were computed only for those participants with all data available at both time points ($n = 356$). SNAP–2 = Schedule for Nonadaptive and Adaptive Personality—2. DIPD-IV = Dimensional Interview for *DSM–IV* Personality Disorders.
[a] Scale correlations were transformed to *z* scores, averaged, and transformed back to correlations.
* $p < .05$. ** $p < .01$. *** $p < .001$.

obsessive-compulsive, borderline, avoidant, histrionic, and narcissistic PDs. Decreases for all other PDs were nonsignificant.

Stability of Categorical Diagnoses

Table 2 presents the rank-order stability values for categorical diagnoses provided by each instrument at both time points. These kappa coefficients ranged from .18 (obsessive-compulsive) to .49 (paranoid), with a median value of .43 for the SNAP–2 and from .12 (narcissistic) to .60 (antisocial), with a median value of .38 for the DIPD-IV. In addition, the 95% confidence intervals around these kappas are presented within Table 2. The confidence intervals for the SNAP–2 and the DIPD-IV overlapped for all 10 PDs, indicating that the kappa values were not meaningfully different across methods.

Finally, Table 2 provides the percentage of individuals meeting each categorical diagnosis at baseline and 2 years, as well as the difference between these two values. This indicates the population level change in the diagnoses for both the SNAP–2 and the DIPD-IV. These percentages demonstrate that although avoidant, borderline, obsessive-compulsive, and schizotypal PDs predominated, all 10 *DSM–IV* PD constructs were represented. Because several prevalence rates were below 5%, however, the kappa should be interpreted very cautiously. Diagnostic rates decreased across the board but were largest for those PDs with the highest initial prevalence. The mean and median of the diagnostic prevalence rate differences are presented at the bottom of the table and suggest that change was relatively consistent across the two instruments, with perhaps a slightly greater decrease for the DIPD-IV.

Discussion

Whereas most previous studies on the longitudinal assessment of PDs have used scores from semistructured interviews, the current study investigated temporal stability of scores from a self-report questionnaire. We observed a mean value of .69 for the rank-order stability for the SNAP-2's dimensional scores over a 2-year period. These findings over 2 years closely resemble the value Melley et al. (2002) obtained for the SNAP over a 9-month interval. These findings also converge with those previously reported by Lenzenweger (1999) in his comprehensive analysis of *DSM–III–R*-based PD dimensions in college students. Thus, the rank-order stability of self-reported PD scores appears no lower in our clinical sample than among undergraduates.

The primary and novel findings of interest from the current study concern the direct comparison of the rank-order and mean-level stability of the PD scores generated via the self-report questionnaire and those from a semistructured interview. It is interesting that the current findings indicate that differential stability does emerge but depends on whether one adopts a dimensional or categorical scoring approach. Whereas dimensional PD scores from a self-report questionnaire demonstrated higher rank-order and mean-level stability than interviews, this was not true for categorical diagnoses. Specifically, for the dimensional scoring, the rank-order stability for SNAP–2 assessments of paranoid, schizoid, histrionic, narcissistic, dependent, and obsessive-compulsive PDs were significantly higher than for the DIPD-IV. Similarly, the SNAP–2 assessments of borderline, histrionic, narcissistic, avoidant, and obsessive-compulsive PDs evinced a smaller mean-level decrease than the DIPD-IV. There was no PD for which the DIPD-IV demonstrated greater dimensional stability by either metric.

The current study is the first to provide a direct statistical comparison between the stability of a self-report and an interview-based assessment. Its results, however, are comparable with those from previous studies. For example, the mean rank-order stability coefficient across the 10 SNAP–2 PD scales in the current study was .69 and the value for the DIPD-IV was .59. These values strongly resemble those reported by Lenzenweger (1999), who reported a mean rank-order correlation of .71 for self-reported PDs and .57 for an interview measure. The similarity between these findings is all the more remarkable because the studies used

Table 2
Stabilities for Categorical Representations of the Personality Disorders

	Rank order (κ)				Mean level (% with diagnosis)					
					SNAP–2			DIPD		
Personality disorder	SNAP–2	95% CI	DIPD–IV	95% CI	Baseline	2 year	Diff.	Baseline	2 year	Diff.
Paranoid	0.49	[.31, .66]	0.47	[.30, .64]	8.8%	6.2%	−2.6%	8.5%	7.1%	−1.4%
Schizoid	0.44	[.27, .58]	0.17	[−.02, .50]	10.7%	8.8%	−1.9%	2.3%	0.8%	−1.5%
Schizotypal	0.42	[.26, .56]	0.58	[.42, .72]	14.1%	10.2%	−3.9%	11.6%	5.9%	−5.7%
Antisocial	0.46	[.24, .66]	0.60	[.38, .77]	5.6%	4.8%	−0.8%	6.8%	5.9%	−0.9%
Borderline	0.31	[.19, .43]	0.50	[.40, .59]	19.5%	13.0%	−6.5%	32.0%	18.4%	−13.6%
Histrionic	0.45	[.28, .60]	0.21	[−.01, .57]	8.8%	9.3%	0.5%	1.7%	0.8%	−0.9%
Narcissistic	0.36	[.12, .57]	0.12	[−.04, .32]	4.5%	4.2%	−0.3%	4.5%	3.1%	−1.4%
Avoidant	0.44	[.34, .53]	0.49	[.40, .59]	42.7%	36.4%	−6.3%	41.1%	30.2%	−10.9%
Dependent	0.24	[.10, .38]	0.27	[.06, .49]	13.3%	8.2%	−5.1%	5.4%	2.3%	−3.1%
Obsessive Compulsive	0.18	[.03, .32]	0.30	[.20, .39]	14.1%	6.2%	−7.9%	34.8%	18.6%	−16.2%
Mdn	0.43		0.38		12.0%	8.5%	−3.3%	7.7%	5.9%	−2.3%
M	0.38		0.37		14.2%	10.7%	−3.5%	14.9%	9.3%	−5.6%

Note. Values presented were computed only for those subjects with all data available at both time points ($n = 356$). SNAP–2 = Schedule for Nonadaptive and Adaptive Personality—2; DIPD-IV = Diagnostic Interview for *DSM–IV* Personality Disorders; CI = confidence interval; Diff. = difference.

different instruments and even assessed PD constructs from different versions of the diagnostic manual (*DSM–III–R* vs. *DSM–IV*). Additionally, the results are consistent with those Trull and Goodwin (1993) reported in examining the temporal stability of *DSM–III–R* PD scores in 44 psychiatric outpatients over 6 months. Trull and Goodwin reported the mean stability across the 10 PDs was .61 for an interview measure, whereas two self-report questionnaires obtained values of .75 and .65. Thus, the current results add to converging evidence suggesting greater rank-order stability for self-reported PD scores than those derived from an interview.

The temporal stability of scores from the SNAP–2 and DIPD-IV were also evaluated in terms of the mean-level stability of the dimensional scores. The mean-level analysis indicated meaningful decreases on scores for all 10 PDs assessed by both methods. However, the overall decrease was larger for the DIPD-IV (mean $d = 0.30$) than for the SNAP–2 (mean $d = 0.21$). We are aware of no previous study that provides a useful context for these results, but it again suggests that self-report questionnaires might be less prone to change across time than semistructured interviews.

This is the first longitudinal study to examine the stability of categorical diagnoses assigned by a self-report questionnaire, as they are typically studied within nonclinical samples with base rates too low for adequate calculations of diagnostic agreement (e.g., Lenzenweger, 1999). Differences between the SNAP–2 and DIPD-IV concerning the stability of the categorical diagnoses could not be tested for significance but did not appear as pronounced as for the dimensional scores. The kappa values were largely similar across the two methods, and the 95% confidence intervals overlapped for all 10 PDs. Additionally, the decreases in diagnostic rates across the 2-year interval did not appear appreciably different, at least when collapsed across the PDs. This suggests that although self-report scores are somewhat more stable, the differences are not as detectable from a categorical viewpoint. It further indicates that many of the important CLPS findings regarding the instability of categorical diagnoses (e.g., Shea et al., 2002; Grilo et al., 2004) remain consistent regardless of the assessment method used.

These findings have important ramifications for the understanding of the stability of PDs relative to other constructs. Although the current results echo previous findings in suggesting that the categorical PD diagnoses are not as stable as indicated in the text of the *DSM–IV–TR* (APA, 2000), they do suggest that dimensional representations evince rather substantial consistency across time. In this regard, it is perhaps helpful to consider these results in the context of the stability of other constructs (e.g., Conley, 1984; Watson, 2004). For example, Reichenberg, Rieckmann, and Harvey (2005) reported that the mean rank-order stability of schizophrenia symptoms was .48 over 2 years, and Larsen, Hartmann, and Nyborg (2008) reported a stability of .85 for general intelligence over even longer intervals. In addition, Roberts and DelVecchio's (2000) meta-analysis of personality stability indicated an overall rank-order coefficient of .64 over nearly seven years for adults aged 30–39 years. In sum, it appears that dimensional PD scores, when assessed via the SNAP–2, are somewhat more stable than even relatively enduring symptoms of other psychiatric disorders (e.g., schizophrenia) but not as stable as intelligence, indicating the possibility of meaningful change over time. Instead, the overall stability of self-reported PD scores appears rather similar to that found for general personality traits.

An even more immediate comparison involves results reported from the CLPS sample. Previous findings from our group have suggested that PDs are less stable than general personality traits of the five-factor model (FFM), because the mean rank-order stability for the PDs and the 30 FFM facets meaningfully differed from one another (Morey et al., 2007). However, that comparison was across methods, as the PDs were assessed via interview and the FFM via self-report. The mean rank-order stability of self-reported PD scores in the current study (.69) appears comparable to the rank-order values for the facets (.67) and domains (.74) of the FFM reported by Morey et al. (2007). In short, the stability of constructs depends on many factors, including content, but also more procedural differences such as the source of the ratings.

Possible Explanations and Future Directions

There are several possible explanations for the finding that dimensional PD scores assessed by a self-report questionnaire have higher stability than those assessed by a semistructured interview. One methodological possibility is that the differences in stability values could simply reflect differences in how the instruments were used. The baseline DIPD-IV interview provided the primary data used to determine CLPS inclusion and PD diagnostic assignment. In contrast, in the current study, we selected a subsample for which SNAP–2 PD scales were not used to determine study inclusion. Thus, the finding of higher self-report stability might simply reflect that the self-reported scores include less systematic measurement error at baseline and thus are less prone to regression to the mean. Three of the PDs for which significant mean-level differences were noted between the interview and self-report method were those oversampled in CLPS (viz., borderline, avoidant, and obsessive-compulsive). Perhaps the greater decrease on the DIPD-IV scores for these PDs might be partially explained by inflated baseline scores. Future research should address this question.

For example, a self-report measure could constitute the sole basis for study inclusion, and researchers could examine stability by both semistructured interview and self-report questionnaire. If the current results solely reflect the instruments' use as inclusion criteria, one would expect such a study to yield reversed results (i.e., interview more stable than self-report). This outcome, though, does not seem likely, as Lenzenweger (1999) selected participants on the basis of a self-report screener, and Trull and Goodwin (1993) used no inclusion measure, yet the self-report questionnaire had greater stability than the semistructured interview in both studies. Nonetheless, future research that directly examines this possibility is crucial, as it has important implications for studying the stability of any construct that is defined by extremity above a given threshold. For example, it might ironically suggest that samples of individuals selected on the basis of extreme scores (e.g., those who meet diagnostic criteria) would be imperfect for studying the temporal stability of that construct, as such samples would artificially exaggerate the decrease on their dimensional scores. Were this the case, it might be preferable to obtain a community sample representative of the population and large enough to ensure an adequate representation of the low base-rate phenomena of interest.

An alternate possibility is that, rather than semistructured interviews underestimating the true stability of PDs, perhaps self-report

questionnaires overestimate their stability. An interviewer might exercise judgment in interpreting an individual's response that increases validity and accuracy. One might speculate that self-report is more prone to finding consistency (i.e., stability) that might not be apparent to others. This consistency might not be a valid indicator of personality pathology, and future research testing this would be useful. However, given the findings of Hopwood et al. (2008) for borderline PD, it seems likely that both methods may be valid but for different aspects of PD. Perhaps descriptions using a self-report questionnaire provide greater validity for the assessment of internal, subjective experiences (e.g., disinterest in close relationships within the schizoid criteria), whereas an interviewer might provide more valid scores for directly observable characteristics that are ego-syntonic (e.g., impressionistic style of speech from histrionic PD). Future research clarifying the validity of these methods would be helpful in creating recommendations for empirically supported assessment and diagnostic practices (e.g., Widiger & Samuel, 2005).

Additional research is needed to better understand the temporal consistency of PDs and arbitrate between the different stability values obtained for self-report questionnaires and semistructured interviews in the current study and previous research. One method would be to examine the relative stability of PD scores provided by other sources. For example, PD ratings provided by a knowledgeable informant have been shown to increment self-report questionnaires (and vice versa) in predicting external criteria (e.g., Clifton, Turkheimer, & Oltmanns, 2005; Klonsky, Oltmanns, & Turkheimer, 2002). Thus, it would be interesting to investigate the temporal stability of informant ratings. Informant descriptions provide an alternative assessment that might be less prone to mood fluctuations and might better assess the observable, interpersonal qualities of PDs. In addition, unlike semistructured interviews, informant reports come from the same person at multiple time points. Although we know of no published research on the subject, the St. Louis Personality and Aging Network (Oltmanns & Gleason, in press) is collecting longitudinal data that include multiple ratings by the same informant. Their stability findings will help address this question.

Additionally, although informant methodology typically relies on ratings by spouses or family members, one might collect PD descriptions from clinical informants. Clinicians could rate their patients using a validated instrument over the course of treatment. This approach is routine in other areas of psychiatry (e.g., Hamilton Rating Scale for Depression; Hamilton, 1960) and could be implemented for PDs. These ratings would likely differ from interview scores because clinicians would complete them on the basis of their experiences with the patient over the course of treatment, rather than responses during a single one- to two-hour interview.

Finally, differences in temporal stability might reflect other distinctions between the composition of self-report questionnaires and semistructured interviews. For example, self-report inventories typically contain more items assessing each PD than do interviews. This stems from obvious practical reasons as the time (and personnel) cost per interview item is greater than the cost for a questionnaire item. Nonetheless, multiple items assessing the nuances of a given construct may yield greater measurement precision and perhaps a superior assessment of the core of each construct (Sanislow et al., 2009). Yet self-report questionnaires need not always contain more items than interviews do. In fact, the Personality Diagnostic Questionnaire (Hyler, 2006) includes only a single item assessing each diagnostic criterion. Notably, the Personality Diagnostic Questionnaire—Revised (Hyler & Rieder, 1987) was one of the two self-report inventories Trull and Goodwin (1993) administered, and its temporal stability coefficient (.65) was somewhat lower than that of the other self-report questionnaire (.75) and only marginally larger than that of the interview (.61). Thus, future research investigating the relative stability of self-report questionnaires and semistructured interviews of equal length would be useful and might also correct for differences in internal consistency. This could be done by using existing self-report inventories that contain fewer items or developing interview measures with more items.

A further relevant point is that interviews and self-report questionnaires might also differ in item content. Many of the diagnostic criteria for the *DSM–IV* PDs are quite behaviorally specific and, consistent with this fact, so are the items on most semistructured interviews. It is possible that the items from the SNAP–2, owing to their inclusion in an instrument designed primarily to assess personality traits, might be less behaviorally specific. It would be possible to investigate such a hypothesis through a content analysis, but this is beyond the scope of the current study. If this was the case, this fact might also contribute to the SNAP–2's greater stability, as general personality styles are likely less prone to change than are specific behavioral manifestations. This could be particularly true as individuals grow older and their life circumstances change (e.g., Tackett, Balsis, Oltmanns, & Krueger, 2009).

Limitations

The current study provides the most comprehensive look to date at the temporal stability of self-reported PD scores within a clinical sample, but it has limitations. The results for the nonstudy PD constructs require cautious interpretation, as the CLPS design recruited individuals who had received diagnoses of at least one of four specific PDs: schizotypal, borderline, avoidant, and obsessive-compulsive. Although this approach offered advantages for studying these four diagnoses, it builds in comorbidity for the other PDs that complicates the potential study of a full range of personality pathology. For example, individuals with clinical diagnoses of narcissistic PD were only included in CLPS if they also had one of the four study diagnoses that was considered primary. Thus, the stability of the primary diagnoses could influence narcissistic PD stability. Nonetheless, a pure case of any particular PD, if it exists at all, is likely the exception rather than the rule. Rates of comorbid PD diagnoses in our study were comparable with those in other studies (e.g., Blashfield, McElroy, Pfohl, & Blum, 1994; Oldham et al., 1995; Stuart et al., 1998).

We selected the subsample for the present analyses because the SNAP–2 was not used to determine study inclusion or assign primary diagnoses. Although this likely increased the range of scores on the 10 PD scales, it did not eliminate the possibility of regression to the mean affecting results. To the degree that the SNAP–2 and DIPD-IV scales measure the same constructs, we would expect that SNAP–2 scores for the four study PDs might also be elevated at baseline. However, this would mean the current results underestimate the stability of PDs from self-report questionnaires. Further, we considered only the data from baseline to 2

years, rather than using the longer term follow-up points, to maximize the available sample size and ultimately statistical power to detect the differences between methods.

Finally, the stability of self-reported PDs was studied using the SNAP–2. Although this instrument contains PD scales that exhibit large convergent correlations with other self-report and interview measures of the *DSM–IV* PDs, it was primarily designed and understood as a measure of a dimensional trait model. Future research that uses other self-report questionnaires that were designed explicitly to assess the PDs would be useful.

Conclusions

The current study provided the first examination of the temporal stability of PD scores from a self-report questionnaire in the CLPS sample. Consistent with previous findings from this and other longitudinal studies, PD scale scores decreased significantly over time. However, the current study also indicates that dimensional scores from a semistructured interview were even less stable than scores from a self-report questionnaire. This finding was consistent with other studies presenting stability results for both assessment methods (e.g., Lenzenweger, 1999; Trull & Goodwin, 1993). It is interesting that the same trend was not observed for categorical representations, as diagnoses assigned by the two methods did not show appreciably different stability. It is not immediately clear why this is the case. However, we hypothesize that it might reflect that categorical scoring equates both methods in terms of the range of possible scores (i.e., 0 or 1). In this way, the finding that categorical diagnoses show similar stability for both methods might arbitrate between potential explanations for the differences noted for dimensional scores. Namely, the creation of categorical diagnoses obviously does not alter the item content of the two instruments or how they were used for study inclusion decisions. Thus, the fact that stability is comparable when assessed categorically suggests that the greater range of possible scores is the most likely explanation for why the SNAP–2 exhibited greater dimensional stability. In any event, it further indicates that perhaps the black–white distinction of categorical diagnoses fails to capture important clinical information and suggests that the dimensional conceptualizations proposed for *DSM–5* (*DSM–5* Personality and Personality Disorders Work Group, 2010) have the potential to improve the diagnosis of personality pathology.

Taken together, the current findings generally support previous findings from CLPS and other longitudinal samples. Namely, when considered categorically, PDs do appear less stable than the *DSM–IV* indicates. However, dimensional scores of the same constructs show temporal consistency that, although lower than cognitive abilities, does exceed relatively enduring psychiatric symptoms and resembles general personality traits, particularly when both are assessed using the same method. Specifically, the current study goes beyond previous work to suggest that the stability of the PD constructs depends at least partially on the method of assessment. It is possible that the differences observed between the self-report questionnaire and semistructured interview reflect the way each method was used in the current study (e.g., as inclusion criteria), the perspective of the person providing the ratings, or more practical considerations such as the number of items within each measure. Future research that continues to investigate the stability and external validity of PD constructs assessed by various methods is highly warranted.

References

American Psychiatric Association. (1987). *Diagnostic and statistical manual of mental disorders* (3rd ed., rev.). Washington, DC: Author.

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

Blashfield, R. K., McElroy, R. A., Pfohl, B., & Blum, N. (1994). Comorbidity and the prototype model. *Clinical Psychology: Science and Practice, 1,* 96–99. doi:10.1111/j.1468-2850.1994.tb00011.x

Chmielewski, M., & Watson, D. (2009). What is being assessed and why it matters: The impact of transient error on trait research. *Journal of Personality and Social Psychology, 97,* 186–202. doi:10.1037/a0015618

Clark, L. A. (1993). *Manual for the schedule for nonadaptive and adaptive personality.* Minneapolis, MN: University of Minnesota Press.

Clark, L. A. (2005). Stability and change in personality pathology: Revelations of three longitudinal studies. *Journal of Personality Disorders, 19,* 524–532. doi:10.1521/pedi.2005.19.5.524

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, MN: University of Minnesota Press.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7,* 309–319. doi:10.1037/1040-3590.7.3.309

Clifton, A., Turkheimer, E., & Oltmanns, T. F. (2005). Self- and peer perspectives on pathological personality traits and interpersonal problems. *Psychological Assessment, 17,* 123–131. doi:10.1037/1040-3590.17.2.123

Cohen, P., Crawford, T. N., Johnson, J. G., & Kasen, S. (2005). The Children in the Community Study of developmental course of personality disorder. *Journal of Personality Disorders, 19,* 466–486. doi:10.1521/pedi.2005.19.5.466

Conley, J. J. (1984). The hierarchy of consistency: A review and model of longitudinal findings on adult individual differences in intelligence, personality, and self-opinion. *Personality and Individual Differences, 5,* 11–25. doi:10.1016/0191-8869(84)90133-8

Craig, R. J., & Olson, R. (1998). Stability of the MCMI-III in a substance-abusing inpatient sample. *Psychological Reports, 83,* 1273–1274. doi:10.2466/PR0.83.7.1273-1274

DSM–5 Personality and Personality Disorders Work Group. (2010). *Personality and personality disorders.* Retrieved on February 22, 2010, from http://www.dsm5.org/ProposedRevisions/Pages/PersonalityandPersonalityDisorders.aspx

First, M. B., Spitzer, R. L., Gibbon, M., & Williams, J. B. W. (1996). *Structured Clinical Interview for DSM–IV Axis I Disorders—Patient version (SCID–I).* New York, NY: Biometrics Research Department, New York State Psychiatric Institute.

Grilo, C. M., Shea, M. T., Sanislow, C. A., Skodol, A. E., Gunderson, J. G., Stout, R. L., & McGlashan, T. H. (2004). Two-year stability and change in schizotypal, borderline, avoidant, and obsessive-compulsive personality disorders. *Journal of Consulting and Clinical Psychology, 72,* 767–775.

Gunderson, J. G., Shea, M. T., Skodol, A. E., McGlashan, T. H., Morey, L. C., Stout, R., & Keller, M. B. (2000). The Collaborative Longitudinal Personality Disorders Study: Development, aims, design, and sample characteristics. *Journal of Personality Disorders, 14,* 300–315.

Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery, and Psychiatry, 23,* 56–62.

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.

Hopwood, C. J., & Richards, D. C. S. (2005). Graduate student WAIS–III scoring accuracy is a function of full scale IQ and complexity of examiner tasks. *Assessment, 12,* 445–454.

Hyler, S. E. (2006). *PDQ–4 Personality Questionnaire.* New York, NY: Human Informatics.

Hyler, S., & Rieder, R. (1987). *PDQ–R Personality Questionnaire.* New York, NY: New York State Psychiatric Institute.

Johnson, J. G., Cohen, P., Kasen, S., Skodol, A. E., Hamagami, F., & Brook, J. S. (2000). Age-related change in personality disorder trait levels between early adolescence and adulthood: A community-based longitudinal investigation. *Acta Psychiatrica Scandinavica, 102,* 265–275.

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,* 300–311.

Larsen, L., Hartmann, P., & Nyborg, H. (2008). The stability of general intelligence from early adulthood to middle-age. *Intelligence, 36,* 29–34. doi:10.1016/j.intell.2007.01.001

Lenzenweger, M. F. (1999). Stability and change in personality disorder features: The Longitudinal Study of Personality Disorders. *Archives of General Psychiatry, 56,* 1009–1015.

Lenzenweger, M. F. (2006). The longitudinal study of personality disorders: History, design, considerations, and initial findings. *Journal of Personality Disorders, 20,* 645–670.

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

Loranger, A. W. (1999). *International personality disorder examination: DSM–IV and ICD-10 interviews.* Odessa, FL: Psychological Assessment Resources.

McDermut, W., & Zimmerman, M. (2005). Assessment instruments and standardized evaluation. In J. Oldham, A. Skodol, & D. Bender (Eds.), *The American Psychiatric Publishing textbook of personality disorders* (pp. 89–101). Washington, DC: American Psychiatric.

McGlashan, T. H., Grilo, C. M., Skodol, A. E., Gunderson, J. G., Shea, M. T., Morey, L. C., & Stout, R. L. (2000). The collaborative longitudinal personality disorders study: Baseline Axis I/II and II/II diagnostic co-occurrence. *Acta Psychiatrica Scandinavica, 102,* 256–264.

Melley, A. H., Oltmanns, T. F., & Turkheimer, E. (2002). The Schedule for Nonadaptive and Adaptive Personality (SNAP): Temporal stability and predictive validity of the diagnostic scales. *Assessment, 9,* 181–187.

Millon, T. (1987). *Millon Clinical Multiaxial Inventory II manual.* Minneapolis, MN: National Computer Systems.

Millon, T. (1994). *Millon Clinical Multiaxial Inventory—III: Manual.* Minneapolis, MN: National Computer Systems.

Morey, L. C., Hopwood, C. J., Gunderson, J. G., Skodol, A. E., Shea, M. T., Yen, S., & McGlashan, T. H. (2007). Comparison of alternative models for personality disorders. *Psychological Medicine, 37,* 983–994.

Morey, L. C., Shea, M. T., Markowitz, M. D., 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.

Morey, L. C., Warner, M. B., Shea, M. T., Gunderson, J. G., Sanislow, C. A., Grilo, C., & McGlashan, T. H. (2003). The representation of four personality disorders by the Schedule for Nonadaptive and Adaptive Personality dimensional model of personality. *Psychological Assessment, 15,* 326–332.

Okada, M., & Oltmanns, T. F. (2009). Comparison of three self-report measures of personality pathology. *Journal of Psychopathology and Behavior Assessment, 31,* 358–367.

Oldham, J. M., Skodol, A. E., Kellman, H. D., Hyler, S. E., Doidge, N., Rosnick, L., & Gallaher, P. E. (1995). Comorbidity of Axis I and Axis II disorders. *American Journal of Psychiatry, 152,* 571–578.

Oltmanns, T. F., & Gleason, M. E. J. (in press). Personality, health, and social adjustment in later life. In L. Cottler (Ed.), *Mental health in public health: The next 100 years.* New York, NY: Oxford University Press.

Ottosson, H., Grann, M., & Kullgren, G. (2000). Test–retest reliability of a self-report questionnaire for *DSM–IV* and ICD-10 personality disorders. *European Journal of Psychological Assessment, 16,* 53–58.

Piersma, H. L. (1989). The stability of the MCMI-II for psychiatric inpatients. *Journal of Clinical Psychology, 45,* 781–785.

Piersma, H. L., & Boes, J. L. (1997). MCMI-III as a treatment outcome measure for psychiatric inpatients. *Journal of Clinical Psychology, 53,* 825–831.

Reichenberg, A., Rieckmann, N., & Harvey, P. D. (2005). Stability in schizophrenia symptoms over time: Findings from the Mount Sinai Pilgrim Psychiatric Center Longitudinal Study. *Journal of Abnormal Psychology, 114,* 363–372. doi:10.1037/0021-843X.114.3.363

Reynolds, S. K., & Clark, L. A. (2001). Predicting dimensions of personality disorder from the domains and facets of the five-factor model. *Journal of Personality, 69,* 199–222.

Roberts, B. W., & DelVecchio, W. F. (2000). The rank-order consistency of personality traits from childhood to old age: A quantitative review of longitudinal studies. *Psychological Bulletin, 126,* 3–25.

Roberts, B. W., Wood, D., & Caspi, A. (2008). The development of personality traits in adulthood. In O. John, R. Robins, & L. Pervin (Eds.), *Handbook of personality: Theory and research* (3rd ed., pp. 375–398). New York, NY: Guilford Press.

Rogers, R. (2001). *Handbook of diagnostic and structured interviewing.* New York, NY: Guilford Press.

Samuel, D. B., Ansell, E. B., Hopwood, C. J., Morey, L. C., Markowitz, J. C., Skodol, A. E., & Grilo, C. M. (2010). The impact of NEO PI–R gender norms on the assessment of personality disorder profiles. *Psychological Assessment, 22,* 539–545. doi:10.1037/a0019580

Sanislow, C. A., Little, T. D., Ansell, E. B., Grilo, C. M., Daversa, M., Markowitz, J. C., . . . McGlashan, T. H. (2009). Ten-year stability and latent structure of the *DSM–IV* schizotypal, borderline, avoidant, and obsessive-compulsive personality disorders. *Journal of Abnormal Psychology, 118,* 507–519. doi:10.1037/a0016478

Shea, M. T., Stout, R., Gunderson, J. G., Morey, L. C., Grilo, C. M., McGlashan, T., & Keller, M. B. (2002). Short-term diagnostic stability of schizotypal, borderline, avoidant, and obsessive-compulsive personality disorders. *American Journal of Psychiatry, 159,* 2036–2041.

Simms, L. J., & Clark, L. A. (2006). The Schedule for Nonadaptive and Adaptive Personality (SNAP): A dimensional measure of traits relevant to personality and personality pathology. In S. Strack (Ed.), *Differentiating normal and abnormal personality* (2nd ed., pp. 431–450). New York, NY: Springer.

Skodol, A. E., Gunderson, J. G., Shea, M. T., McGlashan, T. H., Morey, L. C., Sanislow, C. A., & Stout, R. L. (2005). The Collaborative Longitudinal Personality Disorders Study (CLPS): Overview and implications. *Journal of Personality Disorders, 19,* 487–504.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87,* 245–251.

Stuart, S., Pfohl, B., Battaglia, M., Bellodi, L., Grove, W., & Cadoret, R. (1998). The co-occurrence of *DSM–III–R* personality disorders. *Journal of Personality Disorders, 12,* 302–315.

Tackett, J. L., Balsis, S., Oltmanns, T. F., & Krueger, R. F. (2009). A unifying perspective on personality pathology across the life span: Developmental considerations for the fifth edition of the *Diagnostic and Statistical Manual of Mental Disorders. Development and Psychopathology, 21,* 687–713.

Trull, T. J., & Durrett, C. A. (2005). Categorical and dimensional models of personality disorder. *Annual Review of Clinical Psychology, 1,* 355–380.

Trull, T. J., & Goodwin, A. H. (1993). Relationship between mood changes and the report of personality disorder symptoms. *Journal of Personality Assessment, 61,* 99–111.

Vanbelle, S., & Albert, A. (2008). A bootstrap method for comparing correlated kappa coefficients. *Journal of Statistical Computation and Simulation, 78,* 1009–1015.

Watson, D. (2004). Stability versus change, dependability versus error: Issues in the assessment of personality over time. *Journal of Research in Personality, 38,* 319–350.

Widiger, T. A., & Boyd, S. (2009). Assessing personality disorders. In J. Butcher (Ed.), *Oxford handbook of personality assessment* (3rd ed., pp. 336–363). New York, NY: Oxford University Press.

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

Zanarini, M. C., Frankenburg, F. R., Hennen, J., Reich, D. B., & Silk, K. R. (2005). The McLean Study of Adult Development (MSAD): Overview and implications of the first six years of prospective follow-up. *Journal of Personality Disorders, 19,* 505–523.

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

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.

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

Received May 21, 2010
Revision received November 15, 2010
Accepted November 15, 2010 ■