March, 2008

A comparison of interview and self-report methods for the assessment of borderline personality disorder criteria

Christopher J. Hopwood, *Texas A & M University - College Station*
Leslie C. Morey, *Texas A & M University - College Station*
Maria Orlando Edelen
M. Tracie Shea
Carlos M. Grilo, et al.

Psychological Assessment
2008, Vol. 20, No. 1, 81–85

1040-3590/08/$12.00 DOI: 10.1037/1040-3590.20.1.81

A Comparison of Interview and Self-Report Methods for the Assessment of Borderline Personality Disorder Criteria

Christopher J. Hopwood and Leslie C. Morey
Texas A&M University

Maria Orlando Edelen and M. Tracie Shea
Butler Hospital and Brown Medical School

Carlos M. Grilo, Charles A. Sanislow, and Thomas H. McGlashan
Yale University School of Medicine

Maria T. Daversa, John G. Gunderson, and Mary C. Zanarini
McLean Hospital and Harvard Medical School

John C. Markowitz
New York State Psychiatric Institute and Weill Medical College of Cornell University

Andrew E. Skodol
University of Arizona College of Medicine and Institute for Mental Health Research

Interview methods are widely regarded as the standard for the diagnosis of borderline personality disorder (BPD), whereas self-report methods are considered a time-efficient alternative. However, the relative validity of these methods has not been sufficiently tested. The current study used data from the Collaborative Longitudinal Personality disorder Study to compare diagnostic base rates and the relative validity of interview and self-report methods for assessing functional outcome in BPD. Although self-report yielded higher base rates of criteria endorsement, results did not support the common assumption that diagnostic interviews are more valid than self-reports, but instead indicated the combined use of these methods optimally identifies BPD criteria.

Keywords: borderline personality disorder, self-report, structured interview, predictive validity

Structured interview methods are widely considered the gold standard for the diagnosis of borderline personality disorder (BPD; McDermutt & Zimmerman, 2005) in clinical and research settings for several reasons. First, it is widely assumed that patients with BPD lack the insight and cannot engage in the level of perspective taking necessary to accurately judge their own symptomatology. Furthermore, such patients have been hypothesized to potentially manipulate their self-presentation, and they display certain other response styles that are associated with a distorted presentation of their psychological functioning (e.g., Lloyd, Overall, Kimsey, & Click, 1983).

Empirical research comparing interview with self-report methods has demonstrated fairly low agreement for the assessment of BPD (Zimmerman, 1994; although concordance between BPD interviews also tends to be low; e.g., see Hyler, Skodol, Oldham, Kellman, & Doidge, 1992). Self-report methods generally yield higher diagnostic base rates (Hunt & Andrews, 1992; Hyler, Skodol, Oldham, Kellman, & Rosnick, 1990; Hyler et al., 1992). The discrepancy between these methods is typically attributed to the tendency of questionnaires to "overdiagnose" (McDermutt & Zimmerman, 2005); hence the recommendation that self-report instruments "may be useful as screening measures but are inappropriate for use in making diagnoses" (McDermutt & Zimmerman, 2005, p. 99).

This recommendation is based in part on the a priori presumption of diagnostic superiority of interview methods. Yet comparing diagnostic rates of interview and self-report methods cannot definitively answer the question of relative diagnostic validity, and alternative explanations exist. For example, it is possible that interviews under-pathologize, and that when the methods disagree, self-report methods are more accurate. Another possibility is that each method has greater or lesser validity as a function of the diagnostic item. A third is that both methods present incomplete pictures of symptomatology, and thus should be used in conjunc-

Christopher J. Hopwood and Leslie C. Morey, Department of Psychology, Texas A&M University; Maria Orlando Edelen and M. Tracie Shea, Department of Psychiatry and Human Behavior, Butler Hospital and Brown Medical School; Carlos M. Grilo, Charles A. Sanislow, and Thomas H. McGlashan, Department of Psychiatry, Yale University School of Medicine; Maria T. Daversa, John G. Gunderson, and Mary C. Zanarini, Department of Psychiatry, McLean Hospital and Harvard Medical School; John C. Markowitz, Department of Psychiatry, New York State Psychiatric Institute and Weill Medical College of Cornell University; Andrew E. Skodol, Department of Psychiatry, University of Arizona College of Medicine, and Institute for Mental Health Research.

Mary C. Zanarini is the author of the Diagnostic Interview for Personality Disorders—IV. Portions of this manuscript were presented at the Society for Personality Assessment annual conference. This research was supported by National Institute of Mental Health Grants MH50837, MH50838, MH50839, MH50840, and MH50850; MH01654 to Thomas H. McGlashan; F31MH75543-01A2 to Christopher J. Hopwood; and MH073708 to Charles A. Sanislow.

Correspondence concerning this article should be addressed to Christopher J. Hopwood, who is now at the Department of Psychiatry, Massachusetts General Hospital and Harvard Medical School, 15 Parkman WAC 812, Boston, MA 02114. E-mail: chopwood@partners.org

tion for accurate diagnosis, a conclusion consistent with some recommendations that optimal diagnostic practice requires both methods (Pilkonis, Heape, Ruddy, & Serrao, 1991).

To test these possibilities, we assessed the relative validity of self-report and interview measures of BPD at the level of diagnosis, represented both dimensionally and categorically, and at the level of individual criteria. To evaluate assessment methods in terms of individual criteria, we defined four groups of patients in the Collaborative Longitudinal Personality Disorder Study (CLPS; Gunderson et al., 2000) by comparing interview data with self-report data concerning the presence or absence of each BPD criterion. These four groups were defined as follows: (a) patients for whom both interview and self-report methods agreed that the symptom was positive (*unambiguously present; U/P*), (b) those for whom both methods agreed the symptom was negative (*unambiguously absent; U/A*), (c) those whose interview was positive for a criterion while self-report indicated a negative result (*ambiguous, interview positive; A/IP*), and (d) those whose self-report was positive for a criterion while the interview result was negative (*ambiguous, self-report positive; A/SRP*). Comparisons among these groups with respect to functional outcome can provide important information about the properties of the two assessment methods, as discussed below.

Method

Participants

The current analyses included individuals from the CLPS sample with complete diagnostic and functioning data at baseline ($N = 698$) and at 2-year ($N = 406$) and 5-year ($N = 293$)[1] follow-up. Individuals who were not assessed at 2-year follow-up (i.e., either dropped out of the study or did not return questionnaires) did not differ from completers on our composite measure of functional status, $t(696) = -0.84$, $p > .40$; however, noncompleters did differ from completers at 5-year follow-up, demonstrating lower levels of functioning, $t(696) = 3.28$, $p < .001$. By study design, patients met interview criteria for one of five disorders: BPD, schizotypal PD, obsessive–compulsive PD, avoidant PD, or major depression without PD. At baseline, 248 patients met diagnostic criteria for BPD (5 or more symptoms) according to diagnostic interview, and 319 met criteria according to self-report. Informed consent was obtained after a description of the study, which was approved by local internal review boards. Other features of the sample have been described in detail elsewhere (e.g, McGlashan et al., 2000).

Measures

The Diagnostic Interview for *DSM–IV* Personality Disorders (DIPD-IV; Zanarini, Frankenberg, Sickel, & Yong, 1996) is a structured interview whose content corresponds to diagnostic criteria from the *Diagnostic and Statistical Manual of Mental Disorders—Fourth Edition* (*DSM–IV*; American Psychiatric Association, 1994). The inter-rater and test–retest reliability coefficients for this measure for BPD in a CLPS subsample were .68 and .69, respectively (Zanarini et al., 2000); the internal consistency of the DIPD-IV BPD scale was .81. The Personality Disorder Questionnaire—4 (PDQ–4; Hyler, 1994) was also administered at baseline. Like the DIPD-IV, its item content matches diagnostic criteria in the *DSM–IV*, but is worded to make it appropriate for patient self-report. The internal consistency of this scale in the current sample was .71. The convergence of this widely used measure with other BPD indicators is well established (Hyler et al., 1990), although, as discussed above, researchers often note higher base rates of the disorder for the PDQ relative to interview-based estimates (McDermutt & Zimmerman, 2005).

In comparing interview and self-report methods, it was important to derive a composite criterion of functioning that would not favor one method over the other. Six prospectively administered, methodologically balanced measures of psychosocial functioning (i.e., 3 self-report and 3 interview-derived estimates of functioning across similar domains) were combined to represent a validating outcome variable. Three indicators from the Longitudinal Interval Follow-up Examination (LIFE; Keller et al., 1987) represented social, occupational, and recreational functioning. The LIFE is a commonly used interview in psychiatric research and has shown adequate reliability. Three additional indicators were derived from subscales of the Social Adjustment Scale–Self-Report (SAS-SR; Weissman & Bothwell, 1976) that parallel those of the LIFE in content (i.e., social, occupational, and recreational domains). The SAS-SR is widely used in epidemiological and outcome research and has demonstrated adequate reliability. At each assessment interval, within-sample standardized scores on these six variables were summed to derive an overall estimate of functioning. The internal consistencies for these scores were .69, .76, and .85 at baseline, 2-year, and 5-year intervals, respectively.

Analyses occurred in three stages: the first to compare diagnostic rates of the two methods, the second to compare the concurrent validity of interview and self-report methods for the BPD diagnosis, and the third to compare the validity of these methods for each BPD criterion. The first stage assessed diagnostic agreement for the interview and self-report assessment of each BPD criterion and computed kappa coefficients. The percentage of all patients who received symptom endorsement by each method was also calculated to compare base rates for the nine BPD diagnostic criteria.

In the second stage, interview and self-report methods were compared for the overall BPD diagnosis, conceptualized both dimensionally and categorically. To compare dimensional representations of BPD, we entered symptom sums from each method into hierarchical regression analyses predicting the three functioning composites. The self-report and interview scores were entered both first and second in the prediction of each functioning score across six independent models. The advantage of one method over the other was indicated by the statistical significance of changes in the explained variance.

To compare methods in the assessment of diagnostic categories and individual criteria, we established four groups of patients for each BPD symptom. Patients were assigned for the overall diagnosis and each baseline BPD criterion to one of four groups (U/P, U/A, A/IP, or A/SRP) as described above. Group differences on

[1] Missing data reflect both individuals who dropped out of the study and individuals whose self-report data were incomplete. Sample sizes reflective of attrition only are baseline $N = 733$, 2-year follow-up $N = 667$, and 5-year follow-up $N = 554$.

the psychosocial functioning composite were assessed at intermediate and longer term follow-up, and these functional outcomes were compared by means of one-way analyses of variance (ANOVAs). For each BPD criterion, chi-square tests were conducted to test the possibility that patients differentially dropped out from each group of interest for all nine criteria at both 2- and 5-year follow-up. These analyses suggested significantly differential dropout from baseline to 5-year outcome for four symptoms: self-harm, chronic emptiness, stress-related paranoia, and impulsivity. Thus, the data for the 5-year interval ANOVAs for these four BPD criteria were attrition weighted to correspond to the baseline sample. Significant group differences were followed up with the least significant differences post hoc test.

We assumed that the presence of a BPD symptom would relate positively to psychosocial dysfunction since the overall diagnosis is strongly associated with dysfunction (Skodol et al., 2005). Thus, the critical test of the validity of interview and self-report methods involved the level of dysfunction in each group. Given the assumed salience of convergence across diagnostic methods, it was anticipated that the U/P group would manifest greater dysfunction than the U/A group. Differences in outcome between the ambiguous groups (i.e., A/IP and A/SRP) were interpreted as potential evidence of the differential validity of the methods. For example, if the A/IP group demonstrated significantly greater impairment in functional outcome than the A/SRP group (and, thus, presumably greater similarity to the U/P group), the interview could be considered more valid than self-report. Analyses were conducted for each BPD symptom in turn, for baseline, intermediate (2 years prospective) and longer term (5 years) functioning.

Results and Discussion

Table 1 depicts the diagnostic agreement of self-report and interview methods used in assessing nine BPD criteria. Kappa coefficients across the nine criteria were generally modest (range = .25–.52), indicating limited agreement between the methods. The correlation between dimensional total BPD symp-

Table 1
Diagnostic Agreement of Interview and Self-Report Assessments of Borderline Personality Disorder and Psychosocial Dysfunction Scores

	Diagnostic agreement				Group differences				
	Kappa	% SR +	% Int. +	Year	U/P	U/A	A/IP	A/SRP	*F*
Diagnosis	.53	46	36	0	2.25_{c}	-1.66_{a}	0.04_{b}	0.61_{b}	56.01**
				2	1.50_{b}	-1.57_{a}	0.07_{b}	0.90_{b}	17.07**
				5	0.83_{b}	-2.00_{a}	$-0.52_{a,b}$	-0.17_{b}	8.44**
Abandonment	.28	43	27	0	1.85_{b}	-0.82_{a}	0.16_{a}	-0.17_{a}	16.16**
				2	0.54	−0.39	−0.15	−0.09	0.96
				5	0.15	−0.15	0.44	−1.19	2.91
Unstable	.33	28	38	0	1.76_{c}	-1.19_{a}	$0.88_{b,c}$	0.38_{b}	23.67**
				2	1.45_{b}	-0.99_{a}	0.21_{b}	0.72_{b}	7.33**
				5	0.25_{b}	-1.75_{a}	0.06_{b}	$-0.47_{a,b}$	4.26*
Identity	.25	68	32	0	1.23_{c}	-1.37_{a}	-1.75_{a}	0.14_{b}	18.06**
				2	1.27_{c}	-1.63_{a}	$-0.96_{a,b}$	0.03_{b}	10.49**
				5	0.44_{b}	-1.67_{a}	$-1.55_{a,b}$	-1.28_{a}	4.09*
Self-harm	.52	41	25	0	1.62_{c}	-1.01_{a}	$2.23_{b,c}$	$0.68_{a,b}$	23.46**
				2	1.46_{b}	-1.13_{a}	1.52_{b}	0.64_{b}	11.42**
				5	1.02_{a}	-1.69_{b}	$-1.61_{a,b}$	$-0.46_{a,b}$	6.60**
Moodiness	.37	70	52	0	1.22_{c}	-1.89_{a}	-0.31_{b}	-0.56_{b}	20.08**
				2	0.81_{c}	-1.96_{a}	$0.83_{a,b}$	$-0.14_{b,c}$	9.87**
				5	−0.14	−1.81	−0.81	−1.49	3.00
Emptiness	.33	48	43	0	2.26_{c}	-1.97_{a}	-0.14_{b}	0.59_{b}	59.87**
				2	1.80_{b}	-1.86_{a}	-0.97_{a}	0.95_{b}	22.95**
				5	0.21	−1.68	−0.57	−0.54	3.12
Anger	.41	46	46	0	1.40_{c}	-1.49_{a}	1.16_{c}	-0.44_{b}	30.51**
				2	1.25_{c}	-1.39_{a}	-0.19_{b}	$0.23_{b,c}$	11.12**
				5	0.62_{c}	-2.33_{a}	$-0.03_{b,c}$	$-1.32_{a,b}$	9.59**
Paranoia	.34	41	31	0	2.26_{c}	-1.27_{a}	0.69_{b}	0.29_{b}	34.15**
				2	1.46_{c}	-1.09_{a}	$-0.13_{a,b}$	0.28_{b}	8.95**
				5	0.45_{b}	-2.07_{a}	0.13_{b}	0.45_{b}	8.84**
Impulsivity	.31	63	46	0	1.22_{c}	-1.42_{a}	0.77_{c}	-0.59_{b}	21.92**
				2	0.64_{c}	-0.96_{a}	$0.70_{b,c}$	$-0.51_{a,b}$	4.25*
				5	0.04	−1.61	−0.32	−1.52	3.13

Note. Subscript letters indicate significantly different means according to the least significant differences test (higher scores indicate greater dysfunction, equivalent scores indicate lack of significant mean difference). At Interval 2, four analyses (self-harm, emptiness, paranoia, and impulsivity) used baseline-weighted sample sizes. SR = self-report; Int. = Interview; % SR + and % Int. + refer to the percentage of patients classified as having the given symptom according to each method at baseline. Year refers to follow-up interval (0 = baseline). U/P = both methods agree symptom is present; U/A = both methods agree symptom is absent; A/IP = Interview rated symptom as present but self-report rated symptom as absent; A/SR = self-report rated symptom as present but interview rated symptom as absent.
* $p < .01$, ** $p < .001$.

tom counts ($r = .67$) and kappa coefficient for overall categorical diagnosis ($\kappa = .53$) also suggested modest convergence. Consistent with previous research, self-report yielded higher rates of BPD symptoms than interview, with the two exceptions of the unstable relationships and anger criteria.

Hierarchical regression analyses indicated that interview and self-report dimensional representations of BPD incremented one another in the prediction of functioning composites in four of six models. The interview score incremented self-report at baseline ($\Delta R^2 = .03$, $p < .001$) and 5 years ($\Delta R^2 = .02$, $p < .01$) but not at 2 years, and the self-report score incremented the interview at baseline ($\Delta R^2 = .04$, $p < .001$) and 2 years ($\Delta R^2 = .06$, $p < .001$) but not at 5 years. Analysis at the level of categorical diagnosis suggested similar validity for the two methods (Table 1). These findings support the need for further analyses to determine systematic strengths and weaknesses of these methods at the level of individual criteria.

Table 1 provides the sums of standardized functioning scores for each group at each assessment interval. Overall, the "ambiguous" groups did not tend to systematically differ in functional status. The magnitude of dysfunction was greater for the A/IP group in 15 out of 27 symptom-level comparisons with the A/SRP group. For 4 comparisons, this difference was statistically significant, and each method-specific group had greater dysfunction in 2. By contrast, the U/P group had greater dysfunction than the U/A group for all 27 comparisons, with the difference statistically significant in 22 cases (sign test, $p < .001$). These data strongly support the validity of predictions based upon convergence of self-report and interview, with minimal differences in the overall performance of self-report and interview methods where they disagree. The current results support multimethod assessment of BPD, when feasible.

There was, however, some evidence for relative strengths of one method over the other for particular symptoms. Self-report tended to perform better in assessing more experiential symptoms that would presumably require relatively greater clinical inference by an interviewer (e.g., identity disturbance, chronic emptiness), whereas interview tended to perform better in the assessment of more observable or behavioral symptoms (e.g., self-harm, impulsivity). These results suggest that, contrary to the common convention that interviews have greater general validity than self-report assessments of BPD, clinicians might give differential consideration to self-report and interview assessment data depending on the nature of the symptom being described.

Although these results provide noteworthy support for the use of converging lines of evidence across method in psychological assessment, this study examined one particular diagnostic interview and one particular self-report instrument in assessing one particular disorder. Future research directly comparing these and other instruments is needed to understand the impact of methods on the clinical assessment of BPD specifically, as well as for other problem areas (e.g., Grilo, Masheb, & Wilson, 2001). In addition, research with BPD is complicated by the notable instability associated with its symptom presentation and patterns of comorbidity (e.g., McGlashan et al., 2000, 2005). For example, criterion instability may be one factor accounting for the decreasing relation between symptom presence and functioning over time in the current data. Finally, research that differentiates two factors on which these methods differ, the amount of time spent per criterion and the inclusion of a trained rater, would be important to optimize the assessment of BPD.

References

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders, fourth edition (DSM-IV*) Washington DC: American Psychiatric Association.

Grilo, C. M., Masheb, R. M., & Wilson, G. T. (2001). A comparison of different methods for assessing the features of eating disorders in patients with binge eating disorder. *Journal of Consulting and Clinical Psychology, 69,* 317–322.

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

Hunt, C., & Andrews, G. (1992). Measuring personality disorder: The use of self-report questionnaires. *Journal of Personality Disorders, 6,* 125–133.

Hyler, S. E. (1994). *Personality Diagnostic Questionnaire-4 (PDQ–4).* New York, New York State Psychiatric Institute.

Hyler, S. E., Skodol, A. E., Kellman, D., Oldham, J. M., & Rosnick, L. (1990). Validity of Personality Diagnostic Questionnaire-Revised: Comparison with two structured interviews. *American Journal of Psychiatry, 147,* 1043–1048.

Hyler, S. E., Skodol, A. E., Oldham, J. M., Kellman, H. D., & Doidge, N. (1992). Validity of the Personality Diagnostic Questionnaire-Revised: A replication in an outpatient sample. *Comprehensive Psychiatry, 33,* 73–77.

Keller, M. B., Lavori, P. W., Friedman, B., Nielson, E., Endicott, J., McDonald-Scott, P., & Andreason, 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.

Lloyd, C., Overall, J. E., Kimsey, L. R., & Click, M. (1983). A comparison of the MMPI-168 profiles of borderline and nonborderline outpatients. *Journal of Nervous and Mental Disease, 171,* 207–215.

McDermutt, W., & Zimmerman, M. (2005). Assessment instruments and standardized evaluation. In J. M. Oldham, A. E. Skodol, & D. S. Bender (Eds.), *Textbook of personality disorders* (pp. 89–101). Washington DC, American Psychiatric Association Press.

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

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

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,* 1–9.

Skodol, A. E., Gunderson, J. G., McGlashan, T. H., Dyck, I. R., Stout, R. L., Bender, D. S., et al. (2005). Functional impairment in patients with schizotypal, borderline, avoidant, or obsessive-compulsive personality disorder. *American Journal of Psychiatry, 159,* 276–283.

Warner, M. B., Morey, L. C., Finch, J. F., Gunderson, J. G., Skodol, A. E., Sanislow, C. A., et al. (2004). The longitudinal relationship of personality traits and disorders. *Journal of Abnormal Psychology, 113,* 217–227.

Weissman, M., & Bothwell, S. (1976). The assessment of social adjustment by patient self-report. *Archives of General Psychiatry, 33,* 1111–1115.

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

Zanarini, M. C., Skodol, A. E., Bender, D., Nolan, R., Sanislow, C. A., Schaefer, E., et al. (2000). The collaborative longitudinal personality disorders study: Reliability of Axis I and Axis 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 April 24, 2007
Revision received September 20, 2007
Accepted November 21, 2007 ■