Psychiatr Q (2007) 78:15–24
DOI 10.1007/s11126-006-9023-x

ORIGINAL PAPER

The Personality Disorders Institute/Borderline Personality Disorder Research Foundation Randomized Control Trial for Borderline Personality Disorder: Reliability of Axis I and II Diagnoses

Kenneth L. Critchfield, Ph.D. · Kenneth N. Levy, Ph.D. · John F. Clarkin, Ph.D.

Published online: 11 November 2006

Abstract The Personality Disorder Institute/Borderline Personality Disorder Research Foundation randomized control trial (PDI/BPDRF RCT) is a randomized control trial comparing three treatments for borderline personality disorder (BPD). An important issue for any RCT is diagnostic reliability, demonstration of which is necessary to evaluate claims of a treatment's efficacy for a given population. The present paper examines the interrater reliability of Axis I and II disorders in the context of a high base rate of BPD features for participants referred for inclusion in the RCT. Our results indicate good to excellent levels of interrater reliability for all Axis I and II disorders in this context. Assessors were able to reliably diagnose BPD, exclusionary criteria, and comorbid diagnoses. This data is important for comparing findings and sample composition across different studies using similar sampling strategies, especially as treatments are increasingly being developed and tested for BPD.

Keywords Borderline personality disorder · Randomized control trial · Reliability

The Personality Disorder Institute/Borderline Personality Disorder Research Foundation randomized control trial (PDI/BPDRF RCT) is a controlled outcome study for borderline personality disorder (BPD), in which 90 participants were randomized to one of three manualized and monitored, active psychosocial treatment conditions. These treatments are: (1) Transference Focused Psychotherapy (TFP) [1], a treatment for BPD based on object-relational and psychoanalytic

K. L. Critchfield, Ph.D. · K. N. Levy, Ph.D. · J. F. Clarkin, Ph.D.
New York-Presbyterian Hospital, Joan and Sanford I Weill Medical College of Cornell University, New York, NY, USA

K. L. Critchfield, Ph.D. (✉)
Neuropsychiatric Institute, University of Utah, Suite 1648, IRT Clinic, 501 Chipeta Way, Salt Lake City, UT 84108, USA
e-mail: psykc@psych.utah.edu

principles first applied to BPD by Kernberg [2] and notable for its particular emphasis on frequent interpretation of dynamics manifest in the ongoing therapeutic relationship, (2) Dialectical Behavior Therapy (DBT) [3], a treatment for BPD having good evidence of efficacy [4] and that emphasizes a balance between acceptance and change in its combination of cognitive-behavioral and Zen principles, and (3) Supportive psychotherapy [5], another object-relational and psychoanalytically based treatment for BPD which, in contrast to TFP, eschews transference interpretation and places primary emphasis on development of a collaborative engagement with the patient to foster identity development. Patients received medication according to the treatment algorithm developed by Soloff [6], if clinically indicated. The focus of this paper is on one important aspect of the study, namely the reliability of the diagnostic procedures used to select subjects for randomization to the above treatments. These findings have broader relevance than just to this one study however, as will be described below.

An important issue for any RCT is reliability of the diagnoses used to define the sample. Reliability is a necessary, if not sufficient, condition for validity and constrains its upper limit [7]. Conversely, diagnostic unreliability contributes to error variance and inhibits a study's ability to detect relevant effects. The type of reliability that must be demonstrated in a given study differs depending on the questions addressed by that study and the relevant hypothesized sources of error, the degree to which psychometric properties of the measures have been established in similar contexts, and the nature of the constructs under investigation.

Our primary concern in this paper is to provide detail about the degree of success with which appropriate subjects were identified for inclusion in our RCT based primarily on the presence of BPD and the absence of exclusionary criteria on Axis I (to be described later). This report thus focuses specifically on interrater reliability and diagnostic agreement on Axis I and II of the DSM-IV for two well developed and validated, semi-structured interviews. The International Personality Disorders Examination (IPDE) [8] was used to diagnose BPD as well as assess for presence of co-occurring Axis II features. Rule out criteria and comorbid features on Axis I were assessed using the DSM-IV version of the Schedule for Clinical Interviews and Diagnosis for Axis I, Patient Version (SCID-I/P) [9]. Both the IPDE and SCID-I/P have been previously demonstrated to have good psychometric properties in clinical settings, including good to excellent interrater and test–retest reliabilities in each of their various iterations [8, 10–12].

Previous work has established the psychometric properties of the SCID-I/P and IPDE in general clinical settings with a wide range and variety of disorders—settings with a quite different set of base rates of disorder than is produced by the sampling strategy typically used in treatment studies focused on a particular disorder. A handful of researchers have done more population-specific work involving groups of patients qualifying for a personality disorder. However, no previous study has examined the performance of these instruments limited to the context of a very high base rate of borderline pathology, the very situation we were faced with as we began the RCT.

BPD is well known for the heterogeneity of problem patterns that fit within its boundaries. Any 5 of 9 criteria are required by the DSM-IV for diagnosis, resulting in 256 possible BPD variants. Heterogeneity is even more apparent through observation that there are 315 possible pairings of these variants that share only 1 criterion in common. This means that two patients who are both correctly diagnosed

as having BPD may bear little resemblance to each other in terms of specific personality features. In addition to this internal heterogeneity, BPD is well known for extensive comorbidity on both Axis I and II, leading to still more potential for divergence in terms of the presenting problems, symptoms, and personality features of any two, correctly diagnosed, patients with BPD. Diagnostic agreement for BPD could thus be quite challenging. However, data regarding such agreement is crucial, especially given the increasing calls to develop and validate efficacious treatments for BPD [13–15]. Such calls are a function of the high severity, chronicity, and distress to patients, families, and professionals that is associated with BPD. Researchers will most likely construct their samples as we, and others (e.g., Linehan et al. [4]), have done, from groups of patients selected not directly from a general clinical population, or even from a personality disordered population, but rather based on clinical referrals that act to pre-screen for borderline pathology. In our experience, this informal pre-screening through the referral process generally serves to increase base rates well over the 15–20% levels often reported in diagnostic validation studies, and eliminating from the referral pool most subjects who show few or no borderline features.

Previous authors have documented the importance of interpreting reliability coefficients and their impact on sample construction specifically in the context of the base rate of the disorder under scrutiny [12, 16, 17]. These authors have noted the potential for dramatic differences in reliability estimates provided by tests when applied to contexts with either high or low base rates of a given disorder. With a high base rate of BPD features, the diagnostic risks generally shift from the threat of too many false positives to the threat of too many false negatives. The good news here is that clinical samples chosen from a high base rate of BPD features are very likely to be comprised of bona fide cases of BPD. However, diagnostic agreement can still be challenged due to the need to identify and exclude subthreshold cases that nevertheless present with some borderline features. Also of relevance is that any conditions that are naturally comorbid with BPD (or whose features are similar to BPD) will show altered base rates in the clinical referral pool for which cases must be selected, a state that may or may not prove problematic for diagnosticians.

Careful attention to sample composition and diagnostic agreement will be crucial for comparing results across different studies. The sample-specific reliability data presented here thus represent an important step toward the broader scientific goal of testing treatment efficacy for BPD by providing a reference point for the sample selection procedure itself. In addition, data will be provided regarding the reliability of other diagnostic variables, assessed in the same context, that may be predictive of differential treatment response [18, 19].

Method

Participants included in the present reliability analysis were referred from primarily clinical sources that were aware of the project goals to provide no-cost treatment of BPD for a full year duration in exchange for participation with the research protocol. Participants provided informed consent to use of the videotaped diagnostic interviews for research purposes even if not selected for the study sample. The overall sample selected for the present sets of analyses (details below) was predominantly female (88%). Participant age ranged from 19 to 50 years (M = 33.87, SD = 8.51).

The sample was predominantly Caucasian (77%), unmarried (43% single, 37% separated or divorced), and educated (52% completed college).

Interviewers assessed for inclusionary and exclusionary criteria as part of the process of more broadly assessing DSM-IV-defined pathology for each participant. Inclusion criteria consisted of qualifying for a diagnosis of BPD and providing informed consent for study requirements including videotaped sessions and completion of a series of assessments scheduled throughout the course of the treatment and into a follow-up period. Participants were excluded if they met criteria for untreated major depression of high severity at the time of assessment, substance dependence,[1] past or present history of frank psychosis (i.e., Schizophrenia, Schizoaffective Disorder, Delusional Disorder, Psychosis NOS), or had a history of Bipolar I Disorder.[2]

A total of 8 raters evaluated referred patients over the course of the study, with at least two raters at any one point in time over the course of the study. Six of the raters had a Ph.D. in psychology, one rater was in doctoral level graduate training in psychology, and another rater had an M.S.W. and many years of clinical experience. Each of the raters had been trained in the use of the SCID-I/P by the principal investigator of the RCT (the third author of this report) in conjunction with standard training tapes and materials available for this instrument. Training in the IPDE was provided directly by the developer of the measure (A. Loranger).

For both interviews, training consisted of both didactic and experiential components. Didactic instruction for each rater extended over the course of several weeks, varying somewhat depending on the experience level of raters and logistical needs of the project. Didactics consisted of presentation and review of training materials. Experiential training was also provided along with direct supervision in conjoint interviews and subsequent detailed reviews of videotape and scoring. For the SCID-I/P, training was considered complete when the principal investigator certified that the rater was sufficiently prepared to begin to interview participants independently. For the IPDE, training was considered complete when the developer of the measure certified that the rater was sufficiently prepared to begin to interview participants independently. Ongoing training and supervision was conducted throughout the study in order to prevent rater drift. Weekly training meetings were held in which diagnostic interviews were presented, differences discussed, and consensus rules explicated for application with future ratings.

Participants were randomly selected for reliability analysis. SCID-I/P and IPDE interviews were recorded on videotape and then scored by the interviewer. A separate rater, blind to the interviewer ratings, then reviewed and rated the tapes separately. Overall, 31 pairs of comparisons were made on the SCID-I/P, 23 pairs of comparisons for the full IPDE interview, and an additional 23 pairs of comparisons made solely with the BPD diagnostic criteria of the IPDE.

The kappa statistic was used to calculate interrater reliability for individual categorical diagnoses on Axis I and II. Diagnoses with an interviewer-observed base rate of less than 5% were reported, but identified as unstable [16, 20] in order to provide data in key areas of concern, while also facilitating comparison with prior

[1] Subjects could meet criteria for past major depression or past substance dependence. Subjects could be re-assessed for inclusion in the study if they received treatment for these problems prior to returning.

[2] Low IQ (below 85) was also a rule out criterion. It is not detailed in this report because a separate interview was used for its assessment.

research [8, 21, 22]. A summary kappa was calculated separately for Axis I and Axis II using a weighting procedure to adjust for disorder base rates; see First et al. [23] or Loranger et al. [24] for details of this procedure. Intraclass correlation coefficients, (ICC[1,1] as described by Shrout and Fleiss [25]) for single raters based on random pairings were calculated for the number of DSM-IV Axis II criteria met on the IPDE as well as ratings for the IPDE dimensional scores.

Results

Table 1 details the reliability for Axis I disorders along with results from a previously-published report [22], selected for comparison because it also contains diagnostic agreement data for use of the SCID-I/P for DSM-IV in the context of a personality disordered sample. As can be seen, kappa ranged from 0.59 (anxiety disorder) to 1.00 (substance dependence). These values are in the good to excellent range and are comparable to the other values reported in Table 1. They are also similar to values obtained for the SCID-I/P in other research and training contexts [12, 26, 27]. Overall kappa for Axis I, adjusted for base rates, was 0.77, a level of reliability that can be characterized in the good to excellent range [20]. However, such designations have little meaning outside of a particular research context. For the purposes of our RCT, we concluded from these data that we had met the goal of reliably assessing DSM-based rule out criteria and comorbid features for our participants on Axis I.

For non-BPD Axis II disorders (see Table 2), kappa ranged from 0.70 (Dependent personality disorder) to 1.00 (Paranoid, Antisocial, and Obsessive–Compulsive personality disorders). The base rate for BPD was 76%, which is substantially higher than in instrument validation work involving personality disorder diagnosis, which is usually in the 15–20% range. The current base rate is not so high as to suggest a

Table 1 DSM-IV Axis I base rates and diagnostic agreement with comparison to previous data from a similar setting

Diagnosis	Present study (N = 31)		Zanarini et al. [22] (N = 84)	
	Base rate	Kappa	Base rate	Kappa
Psychotic disorders	0.15	0.84	–	–
Bipolar I	(0.03)	(0.65)	–	–
Bipolar affective disorders	0.10	0.78	–	–
Unipolar affective disorders	0.55	0.74	0.23–0.29	0.79
Anxiety disorders	0.27	0.59	0.10–0.33	0.69
ETOH/Substance abuse	0.39	0.68	0.37	1.00
ETOH/Substance dependence	0.23	1.00	–	–
Eating disorders	0.24	0.90	0.06	0.77
Overall weighted value		0.77		0.85

Present study base rates relied on the interviewer's diagnosis. In order to facilitate cross-study comparison, kappa values were averaged for anxiety and unipolar affective disorders reported by Zanarini et al., with weights applied according to base rates for each category of disorder. In some cases, base rates are reported as ranges due to lack of information regarding rates of comorbidity. The midpoint of the range was used for weighting overall kappa. ''–'' indicates data that is either unavailable or was not reported due to low base rate. Tabled data are adapted with permission from the copyright holder (22, p. 297), Guilford Press, see reference list for details

Table 2 DSM-IV Axis II base rates and diagnostic agreement using the IPDE for a sample of subjects referred to the PDI/BPDRF RCT

Diagnosis	Base rate	% Agreement	Kappa
Paranoid	0.09	100	1.0
Schizoid	(0.00)	(100)	–
Schizotypal	(0.00)	(100)	–
Antisocial	(0.04)	(100)	(1.0)
Borderline	0.76	87	0.64
Histrionic	(0.00)	(100)	–
Narcissistic	0.09	91	–
Avoidant	(0.00)	(91)	–
Dependent	0.22	91	0.70
Obsessive–Compulsive	(.04)	(100)	(1.0)
Overall weighted values		90	0.70

$N = 46$ for BPD. $N = 23$ for other Axis II diagnoses. Base rates are derived from interviewer diagnosis and expressed as a proportion of the reliability sample. Values in parentheses are unstable due to low base rate of the disorder (<0.05). ''–'' indicates kappa cannot be calculated for a given disagreement matrix

problematic level of homogeneity for computing statistics such as kappa, and because is neither too low nor too high actually enhances confidence that ''true positives'' were selected for inclusion in the study. Nevertheless, the kappa produced for BPD diagnosis was only 0.64. There is a rather wide range of similar values published in work involving BPD where kappa has been found to range from 0.40 to 0.96 [17, 28]. The value we obtained is lower than in many of these previous studies, but is still in the fair to good range, and acceptable for our purposes.

Inspection of the cases where disagreements occurred revealed scores that were almost uniformly near the diagnostic threshold such that a single criterion difference between raters (often only differing between ''present'' and ''subthreshold'') would make the difference between diagnostic agreement and disagreement. We take this to suggest that the lower-than-expected reliability reflects difficulty in discriminating subthreshold from full BPD cases. However, the combination of a high base rate of BPD, coupled with the observed level of reliability leads to a strong conclusion of success in constructing a treatment sample homogeneous for BPD diagnosis in the RCT.

Combined kappa for all Axis II diagnoses, adjusting for base rates, was 0.70, placing our overall Axis II diagnostic classifications in the ''good'' range, using Fleiss' [20] designations. This value is somewhat tentative, however, given the low base rates of cases reaching full diagnostic threshold for many non-BPD Axis II disorders in this sample.

Although there were low base rates for diagnosis of many of the Axis II disorders, substantial levels of Axis II features were still present beyond just those associated with BPD. The IPDE provides for dimensional measurement of these features and allows for a test of interrater reliability on the degree of presence of these features rather than the simple yes/no of categorical diagnosis. Intraclass correlation coefficients, shown in Table 3,[3] ranged from 0.67 (Schizotypal) to 0.93 (Paranoid) for dimensional scores (average ICC[1,1] = 0.82). Interrater reliability for dimensions

[3] Difference in sample size between Tables 2 and 3 reflects a clerical error that resulted in loss of dimensional data for 3 subjects who had been rated.

Table 3 Reliability of Axis II dimensional ratings for each personality disorder compared with previous work

Diagnosis	Present study	Maffei et al. [21]	Zanarini et al. [22]	Loranger et al. [29]	Zanarini and Frankenburg [30]
Paranoid	0.93	0.91	0.86	0.97	0.88
Schizoid	0.90	0.93	0.69	0.84	0.55
Schizotypal	0.67	0.94	0.91	0.98	0.70
Antisocial	0.77	0.98	0.97	0.97	0.95
Borderline	0.86	0.95	0.90	0.98	0.96
Histrionic	0.70	0.95	0.83	0.97	0.77
Narcissistic	0.69	0.95	0.88	0.94	0.86
Avoidant	0.86	0.96	0.79	0.97	0.81
Dependent	0.92	0.94	0.87	0.99	0.88
Compulsive	0.90	0.93	0.85	0.96	0.85

Tabled values are ICCs or comparable correlation coefficients. Loranger et al. ([29], p. 10) used the IPDE. Maffei et al. ([21], p. 282) used the SCID-II 2.0. Zanarini et al. ([22], p. 296) reports on the DIPD-IV and Zanarini and Frankenburg ([30], p. 372) used the DIPD-R. These values are reprinted with permission of the respective copyright holders (Elsevier for Zanarini and Frankenburg; Guilford Press for the others, see reference list for details). For the present sample: $N = 43$ for BPD, $N = 20$ for other Axis II diagnoses

was notably higher than for categorical diagnosis. ICC values all fell in the good to excellent range, but were lower than in other studies using semi-structured interviews for Axis II. This was especially notable for disorders traditionally seen as overlapping with BPD, namely the Cluster B disorders and Schizotypal personality.

Discussion

We set out to assess the interrater reliability of our diagnostic judgments in the PDI/BPDRF RCT. Our findings indicate that the assessors were able to reliably diagnose both Axis I and II disorders. Of particular importance to our study, assessors reliably identified BPD and the exclusion diagnoses (psychotic disorders, bipolar I disorder, and alcohol and substance dependence).

The generally high levels of reliability we found with the SCID-I/P and IPDE are consistent with previous research using these measures [8, 11, 26, 27]. Our findings are also consistent with previous interrater reliability studies of DSM-III, DSM-III-R, and DSM-IV criteria sets using other measures of Axis II [21, 28, 30–37]. The finding of a generally high level of reliability occurred despite a much higher presence of borderline pathology than has been observed in other samples. Slight reduction in our reliability levels compared to other studies may be a function of the limitation in range associated with having many cases with subthreshold levels of BPD features, and few cases with little or no presence of such features.

Somewhat lower dimensional reliabilities were also observed for disorders that share conceptual overlap and relatively more comorbidity with BPD. These were Schizotypal personality and the Cluster B disorders; Antisocial, Histrionic, and Narcissistic. It is possible that the lowered reliabilities were related to feature overlap between these disorders, posing difficulties to raters that paralleled those of distinguishing between full and subthreshold BPD cases.

We found that the dimensional ratings from the IPDE exhibited higher levels of interrater reliability than did the categorical ratings. This finding is consistent with previous studies that have compared Axis II pathology measured dimensionally versus categorically [21, 31, 38, 39]. This finding has relevance to the dimensions versus categories debate in the personality disorder literature [40]. A dimensional perspective would predict a lower level of reliability for categorical distinctions because of the introduction of error variance related to use of an arbitrary cutoff score along an otherwise smooth continuum. If the constructs were truly categorical, an optimal cutoff value could be found for which reliability would not be very different across dimensional and categorical ratings. Our findings, while not conclusive on the subject, lend support to the assertion that personality disorder reflects a set of dimensional constructs. This conclusion is tentative, however, given the polythetic nature and multiple conceptual domains of the symptoms and behaviors that lead to personality disorder diagnosis, each with their own measurement quandaries. Personality disorder categories, whether ultimately dimensional or taxonic, are most likely to arise out of the combination of multiple underlying constructs rather than conforming to a unitary, internally consistent, structure. Meehl and colleagues have outlined measurement models and strategies for testing taxonicity under such conditions [41, 42]. Future research along these lines, such as that conducted by Lenzenweger [43] with Schizotypal personality, is needed to untangle the relevant subtleties.

A major strength of our study includes use of a largely BPD sample from which to view the issue of reliability. This rating context will become increasingly relevant for future comparison of BPD treatment studies. However, there are limitations to our design deriving primarily from the fact that our major goal was to construct a treatment sample, rather than to measure reliability per se. For example, the sample was relatively small, prohibiting finer degrees of resolution regarding individual criteria. Low base rates of many disorders led to likely instability of many kappa values. Use of videotaped rather than conjoint or separate interviews limited our consideration of alternate sources of variance that may arise from interviewing style, state effects, etc. Finally, all raters did not assess all subjects, and so estimates of variability due to individual raters, and any interactions of raters by cases could not be pursued.

Conclusion

The present results indicate good to excellent levels of interrater reliability for all Axis I and II disorders in the context of the high level of borderline pathology (and related restriction of range of this pathology) referred to our study. Assessors were able to reliably diagnose BPD, exclusionary criteria, and comorbid diagnoses. We are thus confident in having been able to create a subject pool for randomization to treatment that was reliably homogeneous for BPD diagnosis and that did not show evidence of the rule out criteria. We are also reasonably confident in our ability to characterize the sample in terms of comorbid conditions on Axis I and II that may serve as important moderators of treatment efficacy in the overall RCT. These results should help provide a benchmark for reliability results reported in similar settings involving a sampling method that informally pre-screens for diagnosis, as is often the case in treatment trials.

Acknowledgments This research was supported by a grant from the Borderline Personality Disorder Research Foundation to Drs. Otto Kernberg and John Clarkin. The authors wish to thank Jack Barchas, M.D. for institutional support and to acknowledge the technical assistance of Catherine Eubanks-Carter, Jill C. Delaney, M.S.W., Pamela E. Foelsch, Ph.D., Simone Hoermann, Ph.D., Maya Kirschner, Ph.D., and Joel McClough, Ph.D. for their help in conducting assessments, James Hull, Ph.D. for organizing and maintaining the data. We also acknowledge the consultation of Armand Loranger, Ph.D. The authors wish to thank members of the Personality Disorders Institute. Finally, we would like to thank the patients for their participation in the project.

References

1. Clarkin JF, Yeomans FE, Kernberg OF: Psychotherapy for Borderline Personality. New York: Wiley, 1999.
2. Kernberg OF: A psychoanalytic theory of personality disorders. In: Clarkin JF, Lenzenweger MF (Eds) Major Theories of Personality Disorder. New York: Guilford, 1996.
3. Linehan MM: Cognitive Behavioral Treatment of Borderline Personality Disorder. New York: Guilford, 1993.
4. Linehan MM, Armstrong HE, Suarez A, et al.: Cognitive behavioral treatment of chronically parasuicidal borderline patients. Archives of General Psychiatry 48:1060–1064, 1991.
5. Appelbaum AH, Carsky M: Supportive Therapy for Borderline Patients. Unpublished manuscript. Author, 2003.
6. Soloff PH: Psychopharmacology of borderline personality disorder. Psychiatric Clinics of North America 23:169–192, 2000.
7. Nunnally JC: Psychometric Theory. New York: McGraw Hill, 1967.
8. Loranger AW: International Personality Disorder Examination (IPDE) Manual. Odessa, FL: Psychological Assessment Resources, 1999.
9. First MB, Spitzer RL, Gibbon M, et al.: Structured Clinical Interview for DSM IV Axis I Disorders, Patient Edition (SCID I/P, Version 2.0). New York: NY, Biometrics Research Department, New York State Psychiatric Institute, 1996.
10. Becker DF, Grilo CM, Edell WS, et al.: Diagnostic efficiency of borderline personality disorder criteria in hospitalized adolescents: Comparison with hospitalized adults. American Journal of Psychiatry 159:2042–2047, 2002.
11. Pilkonis PA, Heape CL, Ruddy J, et al.: Validity in the diagnosis of personality disorder: The use of the LEAD standard. Psychological Assessment 3:46–54, 1991.
12. Williams JBW, Gibbon M, First MB, et al.: The structured clinical interview for DSM III R (SCID) II. Multi site test retest reliability. Archives of General Psychiatry 49:624–629, 1992.
13. Linehan MM: The empirical basis of Dialectical Behavior Therapy: Development of new treatments versus evaluation of existing treatments. Clinical Psychology: Science and Practice 7:113–119, 2000.
14. Scheel KR: The empirical basis of Dialectical Behavior Therapy: Summary, critique, and implications. Clinical Psychology: Science and Practice 7:68–86, 2000.
15. Widiger TA: The science of Dialectical Behavior Therapy. Clinical Psychology: Science and Practice 7:101–103, 2000.
16. Grove WM, Andreasen NC, McDonald-Scott P, et al.: Reliability studies of psychiatric diagnosis: Theory and practice. Archives of General Psychiatry 38:408–413, 1981.
17. Zimmerman M: Diagnosing personality disorders: A review of issues and research methods. Archives of General Psychiatry 51:225–245, 1994.
18. Clarkin JF, Levy KN: A psychodynamic treatment for severe personality disorders: Issues in treatment development. Psychoanalytic Inquiry 23:248–267, 2003.
19. Clarkin JF, Levy KN, Lenzenweger M, et al.: The Personality Disorders Institute/Borderline Personality Disorder Research Foundation randomized control trial for borderline personality disorder: Rationale, methods, and patient characteristics. Journal of Personality Disorders 18:51–71, 2004.
20. Fleiss JL: Statistical Methods for Rates and Proportions. 2nd Ed. New York: Wiley, 1981.
21. Maffei C, Fossati A, Agostoni I, et al.: Interrater reliability and internal consistency of the Structured Clinical Interview for DSM IV Axis II Personality Disorders (SCID II), version 2.0. Journal of Personality Disorders 11:279–284, 1997.

22. Zanarini MC, Skodol AE, Bender D, et al.: The collaborative longitudinal personality disorders study: II. Reliability of Axis I and II diagnoses. Journal of Personality Disorders 14:291–299, 2000.
23. First MB, Spitzer RL, Gibbon M, et al.: The Structured Clinical Interview for DSM III R Personality Disorders (SCID II) Part II: Multisite test retest reliability study. Journal of Personality Disorders 9:92–104, 1995.
24. Loranger AW, Janca A, Sartorius N (Eds): Assessment and Diagnosis of Personality Disorders: The ICD 10 International Personality Disorder Examination (IPDE). New York: Cambridge University Press, 1997.
25. Shrout PE, Fleiss JL: Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin 86:420–428, 1979.
26. Skre I, Onstad S, Torgersen S, et al.: High interrater reliability for the Structured Clinical Interview for DSM III R Axis I (SCID I). Acta Psychiatrica Scandinavica 84:167–173, 1991.
27. Ventura J, Liberman RP, Green MF, et al.: Training and quality assurance with the Structured Clinical Interview for DSM IV (SCID I/P). Psychiatry Research 79:163–173, 1998.
28. Farmer RF, Chapman AL: Evaluation of DSM-IV personality disorder criteria as assessed by the structured clinical interview for DSM-IV personality disorders. Comprehensive Psychiatry 43:285–300, 2002.
29. Loranger AW, Susman VL, Oldham JM, et al.: The Personality Disorder Examination: A preliminary report. Journal of Personality Disorders 1:1–13, 1987.
30. Zanarini MC, Frankenburg FR: Attainment and maintenance of reliability of Axis I and II disorders over the course of a longitudinal study. Comprehensive Psychiatry 42:369–374, 2001.
31. Loranger AW, Sartorius N, Andreoli A, et al.: The International Personality Disorder Examination: The World Health Organization/Alcohol, Drug Abuse, and Mental Health Administration international pilot study of personality disorders. Archives of General Psychiatry 51:215–224, 1994.
32. Cornell DG, Silk KR, Ludolph PS, et al.: Test retest reliability of the Diagnostic Interview for Borderlines. Archives of General Psychiatry 40:1307–1310, 1983.
33. Frances A, Clarkin JF, Gilmore M, et al.: Reliability of criteria for borderline personality disorder: A comparison of DSM III and the Diagnostic Interview for Borderline Patients. American Journal of Psychiatry 141:1080–1084, 1984.
34. Hurt SW, Hyler SE, Frances A, et al.: Assessing borderline personality disorder with self report, clinical interview, or semistructured interview. American Journal of Psychiatry 141:1228–1231, 1984.
35. Kroll J, Pyle R, Zander J, et al.: Borderline personality disorder: Interrater reliability of the Diagnostic Interview for Borderlines. Schizophrenia Bulletin 7:269–272, 1981.
36. Shea MT, Stout R, Gunderson J, et al.: Short term diagnostic stability of schizotypal, borderline, avoidant, and obsessive compulsive personality disorder. American Journal of Psychiatry 159:2036–2041, 2002.
37. Stangl D, Phofl B, Zimmerman M, et al.: A structured interview for the DSM III personality disorders. Archives of General Psychiatry 42:591–596, 1985.
38. Dreessen L, Arntz A: Short interval test retest interrater reliability of the Structured Clinical Interview for DSM III R Personality Disorders (SCID II) in outpatients. Journal of Personality Disorders 12:138–148, 1998.
39. Smith TL, Klein MH, Benjamin LS: Validation of the Wisconsin Personality Disorders Inventory IV with the SCID II. Journal of Personality Disorders 17:173–187, 2003.
40. Widiger TA: Personality disorder dimensional models proposed for DSM IV. Journal of Personality Disorders 5:386–398, 1991.
41. Meehl PE: Bootstraps taxometrics: Solving the classification problem in psychopathology. American Psychologist 50:266–275, 1995.
42. Waller NG, Meehl PE: Multivariate taxometric procedures: Distinguishing types from continua. Thousand Oaks, CA: Sage, 1998.
43. Lenzenweger MF: Deeper into the schizotypy taxon: On the robust nature of Maximum Covariance Analysis. Journal of Abnormal Psychology 108:182–187, 1999.