Running Head: Clinicians' Use of PD Diagnoses

A Review of the Agreement between Clinicians' Personality Disorder Diagnoses and those from Other Methods and Sources

Douglas B. Samuel, Ph.D.

Purdue University

In press*; Clinical Psychology: Science and Practice*

Author Note:

Douglas B. Samuel, Department of Psychological Sciences, Purdue University.

The author wishes to thank Tom Widiger for his helpful comments and critiques on an earlier version of this manuscript.

Correspondence concerning this article should be addressed to Douglas B. Samuel, Department of Psychological Sciences, Purdue University, 703 Third Street, West Lafayette, IN 47907.

Email: dbsamuel@purdue.edu

Abstract

This review synthesizes a wide literature on the agreement of treating clinicians' PD diagnoses with each other and their convergence with common research methods. Median interrater reliability between clinicians was moderate when calculated dimensionally (r = .46) or categorically (κ = .40). The agreement between clinicians' diagnoses and those from research methods (e.g., self-report questionnaire) was more modest. Median dimensional agreement across 27 studies ranged from .05 to .36 with an overall median of .23. This overall value was moderated by several factors. First, clinicians' diagnoses agreed more with semistructured interviews than self-report questionnaires. Second, convergence increased slightly when clinicians utilized more systematic diagnostic methods. Results suggest relatively little overlap between PD diagnoses assigned in research versus naturalistic settings.

Keywords: *clinician, diagnosis, personality disorder, systematic, therapist, practice*

A Review of the Agreement between Clinicians' Personality Disorder Diagnoses and those from Other Methods and Sources

Personality Disorders (PDs) trace their history to the very origins of psychiatry, when the French physician Philippe Pinel offered a description of a certain group of mental patients who lacked delusions, hallucinations, or impaired intellectual functioning, and yet displayed consistently maladaptive social behavior. He labeled this group with a category of *manie sans délire*—insanity without delusion (Pinel, 1801, 1962). This early history recognizes the importance of identifying PDs in clinical practice as they are associated with impairment of an individual's functioning that is significant (Gunderson et al., 2011) and even exceeds that of other mental disorders (Noren et al., 2007). PDs are associated with severe, self-reported quality of life impairment that is equivalent to that reported by individuals with other psychiatric and medical conditions such as lung cancer, arthritis, and Parkinson's disease (Soeteman, Verheul, & Busschbach, 2008). A PD diagnosis quadruples all-cause mortality risk, relative to those without a PD (Eaton et al., 2008), and portends poor treatment prognosis for individuals with other psychiatric diagnoses (Bieling et al., 2003; Meier & Barrowclough, 2009; Newton-Howes, Tyrer, & Johnson, 2006). In addition to these problematic impacts on the individual, PDs also pose a significant public health burden in terms of their relations with violence (Fountoulakis, Leucht, & Kaprinis, 2008), substance use (Trull, Jahng, Tomko, Wood, & Sher, 2010), and health service utilization (Bender et al., 2001).

Despite their clear importance for a variety of consequential outcomes, the valid identification of PDs within clinical practice remains complex and imperfect (Westen, 1997; Zimmerman, 1994). Several methods exist for assessing PDs that could be used to inform diagnosis, including semistructured diagnostic interviews and self-report questionnaires

answered by the client (McDermut & Zimmerman, 2005; Widiger & Samuel, 2005b). For a variety of reasons, though, the vast majority of clinicians base PD diagnoses on their unstructured interviews and clinical contacts with patients (Perry, 1992).

Although there are formal instruments and/or methods designed to systematically aggregate clinicians' diagnostic impressions, such as the Shedler-Westen Assessment Procedure (SWAP; Shedler & Westen, 1998) or the LEAD method (Longitudinal, Expert, All Data; Pilkonis, Heape, Ruddy, & Serrao, 1991) they are not used routinely within practice settings. The SWAP is an instrument that was explicitly designed to capture clinicians' ratings and consists of 200 statements relevant to the description of personality pathology (e.g., tends to be passive and unassertive). The format of the SWAP requires clinicians to sort the 200 statements into eight different piles indexing the degree to which each item applies to the individual. It further utilizes a fixed distribution such that 100 of the items must be sorted as irrelevant or inapplicable, while decreasing numbers are sorted in the remaining seven categories, with only eight total items sorted into the most descriptive category. This rating process requires approximately 45-60 minutes for a clinician (Shedler & Westen, 1998) and the resulting profile can be compared to empirical prototypes to arrive at diagnostic scores for each PD.

The LEAD method of diagnosis relies on expert judges (typically a panel of 3 to 5 mental health professionals) who incorporate all sources of data available to arrive at a consensus diagnosis. For example, the panel of judges may review information provided by the client or a significant other collected via a questionnaire or semistructured interview, data within the patient's clinical chart, and/or ratings provided by any research staff who has interacted with the individual. All data is then reviewed at a case conference that allows detailed discussion of the client whereby diagnostic impressions can be shared and debated before ultimately arriving a

consensus rating. The case conference for each patient typically lasts between 1.5 and 2 hours (Pilkonis et al., 1991).

Given the labor-intensive nature of both the SWAP instrument and the LEAD method, it is perhaps not surprising that they are not routinely utilized in practice settings. Thus, the clinical diagnoses provided by treating therapists are most frequently "unstructured" in the sense that they are not derived from any systematic assessment of the PDs or the diagnostic criteria. Rather, they rely on the expertise of the individual clinician to detect and identify personality pathology during their routine clinical interactions. It is, therefore, extremely important to understand the reliability and validity of routine PD diagnoses assigned by treating clinicians.

Agreement between Clinicians

Interrater reliability, the diagnostic agreement between two separate clinicians describing the same patient, is a fundamental component for the valid use of a PD system. After all, if two clinicians cannot agree reasonably on a diagnosis between themselves, then their diagnoses have little chance of relating well with external criteria or predicting response to treatment. Unfortunately, as Kraemer, Kupfer, Clarke, Narrow, and Regier (2012) noted "many books and articles have been written on the methods of evaluation of medical treatments, but little attention has been paid to the evaluation of the quality of diagnoses" (p. 14). Instead, Kraemer and colleagues argued that interrater reliability should be among the central concerns of a diagnostic system and specifically proposed that categorical agreement between .20 - .40 and dimensional correlations between .40 and .60 would be considered acceptable. These thresholds might be considered overly liberal as interrater agreement at these levels would still suggest quite a lot of unshared variance between independent raters. Nonetheless, the current review summarizes the

existing literature on the interrater reliability of PD diagnoses and uses these benchmarks as minimal thresholds in order to quantify the observed agreement between two separate clinicians.

Agreement across Methods and Sources

In contrast to the unstructured diagnoses typically provided by treating clinicians in routine practice, the diagnostic approach within research settings is quite different. PD diagnoses in research are typically generated based on one or more self-report questionnaires and/or semistructured diagnostic interviews that comprehensively assess each criterion for all PDs (Widiger & Samuel, 2005b). The differences between research and clinical procedures raise pressing questions about the degree to which these contrasting methods agree and what this portends for evidence-based practice. Indeed, if the samples of individuals with PDs in research studies (diagnosed primarily via semistructured interviews) have important differences from how those PDs are diagnosed in clinical practice, this may limit or complicate the adoption of empirically supported strategies. Thus, an aim of this review was to summarize the agreement between PD diagnostic ratings provided by treating clinicians and those from other independent methods (i.e., the instruments employed) and/or sources (i.e., the person providing the ratings).

There have been prior meta-analyses or reviews that have examined portions of this question. For example, Achenbach, Krukowski, Dumenci, and Ivanova (2005) reviewed 108 studies that reported agreement between a self-report instrument and some other method of assessing all types of adult psychopathology. They first noted that less than 0.5% of the studies examined had provided an index of cross-informant correlations, suggesting that "relatively little attention has been paid to this problem" of cross-method agreement in psychopathology research (p. 373). Second, they noted that the mean correlation between self-reports and other sources (e.g., peers, clinicians, interviews conducted by researchers) was .45 when the exact same instrument was

used by both parties (i.e., same method, but different source) and .30 when both the sources and methods were different. Although PDs were investigated as a moderator in these analyses, they did not obtain significantly different agreement than other forms of psychopathology. These findings are extremely informative with regard to the general level of agreement of self-report measures with other diagnostic sources, but do not answer the question about the validity of clinicians routine PD diagnoses as Achenbach and colleagues had a decidedly different purpose. While the studies included in that review were very broad in terms of content, they were limited to only those that utilized a self-report measure as the criterion. Thus, the review by Achenbach and his colleagues excluded potentially valuable information regarding the specific agreement of clinicians with other sources, such as semistructured diagnostic interviews.

Meyer and colleagues (2001) provided a thorough review that indexed the cross-method agreement of a wide variety of psychological assessment instruments. A key finding was that alternate methods and sources do provide unique information that is valuable in an assessment or diagnostic context. Within this larger review, Meyer and colleagues (2001) summarized the specific agreement between "self vs. clinician" (p. 148; Table 3) for the DSM PDs and reported a median categorical agreement (kappa) coefficient of .18 and a median dimensional correlation of .33. Nonetheless, Meyer and colleagues employed a rather broad interpretation of the term "clinician" in their review, even classifying researcher-administered semistructured interviews as clinician ratings, again limiting the conclusions that can be drawn about the validity of routine clinical diagnoses of PDs. Not only might this methodological decision inflate convergent values, but it also precludes the comparison of semistructured interviews (administered by researchers) to routine diagnoses by treating clinicians. What would be most informative in this regard is a review that focuses explicitly on the *source* of the ratings, regardless of the method

utilized. Such a review would then index the agreement between the individual's treating clinician and any other source or method.

Widiger and Boyd (2009) conducted a more focused review of this literature, considering 21 studies that reported on the agreement of unstructured clinical interviews and even broke down the results by the individual PDs. Although they reported a number of different values, the most relevant to the current review was that the median dimensional agreement for the individual PDs was .54. However, as those authors noted, this overall value collapsed across a variety of studies that used substantially different methods. For example, the global estimates were derived from studies reporting interrater reliability (e.g., Mellsop, Varghese, Joshua, & Hicks, 1982) as well as those that were within-, rather than across-method. As Widiger and Boyd noted explicitly, the largest effect sizes were from studies where the same clinicians provided both the PD diagnoses and criterion ratings for each patient (Westen & Shedler, 1999). Within their review, Widiger and Boyd differentiated between those studies where the ratings were blind to each other and noted that that median value for studies that did not suffer from criterion contamination was approximately half the magnitude of those that did (i.e., median value of .27 compared to .55). The current review seeks to build upon that particular comparison by including more recently published findings and distinguishing between ratings provided by treating clinicians and those assigned via other methods (i.e., LEAD diagnoses by a research team; Pilkonis et al., 1995). On the basis of the findings of Widiger and Boyd (2009) it is expected that the overall agreement across sources will be more comparable to .27 than .50.

Another relevant comparison point for the present study is the review by Klonsky, Oltmanns, and Turkheimer (2002), which considered 11 studies that reported agreement between PD ratings provided self-report questionnaires and peer informants. The overall median level of categorical

agreement across these self and informant ratings was κ = .14, while dimensional agreement was *r* = .36, suggesting that self-report and informant reports of PD shared commonality, but also had important differences. Thus, one might reasonably hypothesize that the agreement between clinicians (a specific sort of informant) and other methods would approximate this value.

Current Review of the Literature

The current review sought to identify studies that reported the agreement between PD diagnostic ratings provided by treating clinicians in the course of their therapeutic contact and some other diagnostic source. The rationale is that this focus provides the most accurate index of the agreement that can be expected between PD diagnoses in routine clinical settings and those provided by other methods. In this same vein, the current review separates those studies that utilized naturalistic diagnostic methods (i.e., brief and mostly unsystematic assessments) and compares them to others that used more systematic assessment tools to collect ratings from the clinicians (i.e., SWAP). This allows a formal comparison of the potential validity differences when using structured methods versus the typical unstructured ratings.

Studies Reviewed

The present review primarily followed the method utilized by Klonsky et al. (2002) to index the agreement between self-report and informant ratings of PDs. In order to secure a complete list of relevant studies the reference lists of Perry (1992), Achenbach et al (2005), Meyer et al (2001), and Widiger and Boyd (2009) were first consulted. As this current review had a different scope and purpose than those others, each of those reference lists contributed unique studies. After generating an initial list, Web of Science was searched in July 2013 with the following search string: Topic = (Personality disorder* OR axis II) AND Topic = (clinician* OR therapist*) AND Topic=(rating* OR diagnos* OR descri* OR assess* OR report*) AND

Topic = (clinical OR unstructured OR routine OR practice). This search yielded 1,863 results. Specific studies from these lists were excluded for a variety of reasons. For example, Marin-Avellan and colleagues (2005) was excluded because the SCID-II and SWAP-200 were completed by the same researchers and were not based on clinical contact by a treating clinician. Additionally, work by Pilkonis and colleagues (1991, 1995) using the LEAD method was excluded because diagnoses were made by the research team based on the research interviews and not based on therapeutic contact. An examination of these studies yielded 36 manuscripts that met inclusion criteria. This list contained several that were not included in prior reviews, as well some that were published subsequently indicating the novelty and comprehensiveness of the present review.

Methodological Considerations

Many of the studies qualifying for inclusion still had notable methodological differences. For example, fifteen studies considered only a subset of the ten current PDs. This was due in some cases to an a priori focus on specific constructs (e.g., Samuel et al., 2013) and in others to the use of measures that assessed only a subset (Löffler-Stastka et al., 2006). Still other studies had incomplete data as insufficient base rates for some PDs prevented calculations of categorical agreement. An additional four studies did not provide results for individual PDs at all, and instead reported only the agreement for the presence versus absence of any PD diagnosis (e.g., Cantrell & Dana, 1987). Further, due to the range of years this review covers, the specific operationalization of the PDs also varied. Seventeen of the studies concerned PDs as defined by *DSM-IV*, nine used *DSM-III-R* definitions, six *DSM-III*, two *ICD-10*, and two utilized Q-variables which are similar to the DSM-IV PDs, but are specific to the SWAP-200.

There were also differences in the statistics utilized to report agreement across the studies. Although two studies (Allard & Grann, 2000; Hyler, Rieder, Williams, & Spitzer, 1989) reported both categorical and dimensional agreement statistics, most focused on only one or the other. Of the remaining 34 studies, 18 utilized Kappa to index categorical agreement while 16 utilized Pearson correlations to quantify dimensional agreement.

Finally, although all studies had to report PD ratings provided by a treating clinician for inclusion in the review, there was substantial variation in the specific instruments used to collect and aggregate those ratings. Fourteen studies did not use any instrument and clinicians simply assigned one or more categorical diagnoses, such as might be done within a medical chart. The remaining 20 studies employed a variety of measures. These measures were mostly brief rating forms at the level of PD (e.g., Personality Assessment Form; Shea, Glass, Pilkonis, Watkins, & Docherty, 1987), but others were much longer and collected ratings on individual items or diagnostic criteria that were combined to render PD diagnoses.

In addition to the variation in the instruments used to collect clinician ratings, the source providing the criterion ratings varied across studies, as did the instruments they used to provide those ratings. Nine of the studies concerned exclusively the agreement between two independent clinician raters (i.e., interrater reliability), 14 compared clinicians' ratings to those provided by clients via a self-report questionnaire, nine utilized a semistructured interview administered to clients by research personnel, and four considered multiple sources and/or methods.

In short, the methodological variation across the studies prohibited a formal aggregation using meta-analytic methods (e.g., sample-size weighted effect size estimates). Like Klonsky and colleagues (2002) the median values across studies are reported with each study weighted equally. In cases where a single study provided more than one effect size (e.g., two separate

criterion measures; dimensional and categorical agreement statistics), both were included in omnibus calculations, but disaggregated for specific moderator analyses.

Results

Interrater Reliability

A first step was to clarify the agreement between raters for PD diagnoses within the nine studies reporting interrater reliability. Table 1 provides the Kappa values for the diagnosis of any PD (vs. No PD) and they ranged from .35 to .62, with a median of .52. Categorical agreement was quite similar regardless of whether the raters had interviewed the client jointly (.52) or separately (.48). The median kappa agreement for individual PD diagnoses in each study ranged from .23 to .62, with an overall median of .40. When PDs were rated dimensionally, the median Pearson correlations for each study ranged from .34 to .74, with an overall median value of .46. There did appear to be some distinction based on the instrument used to aggregate the clinicians' ratings as the median was .61 for two studies that had used a version of the SWAP, but .43 for those that used other methods.

Cross-method Agreement

Table 2 presents the studies that concerned the agreement of clinicians' ratings with other methods and/or sources. The overall median agreement coefficients across all 27 studies and all methodologies are presented first to summarize the best available estimate of the convergence between clinicians' diagnoses and those from other sources. The overall kappa agreement for the diagnosis of any PD (vs. No PD) ranged from a low of -.07 to a high of .96, with a median of .15. The median categorical agreement for a specific PD diagnosis across all studies ranged from a Kappa of .03 to .65, with a median of .26. When considered dimensionally, the median correlation ranged from .05 to .36 with an overall median of .23.

The studies were then disaggregated by the source of the criterion ratings (i.e., self-report or semistructured interview) to determine if there were differences. Table 2 indicates that there was a notable distinction based on the source of the criterion ratings. The overall median kappa for any PD diagnosis when it was between clinicians' ratings and a self-report questionnaire was .11 (n = 4). The equivalent value when a semistructured interview was used as the criteria was .38 (n = 5). This was similar for the categorical agreement for specific PD diagnoses, as self-report (κ = .08, n = 6) was again lower than semistructured interview (κ = .30, n = 11). This same trend was not nearly noticeable when PDs were rated dimensionally, with a median correlation of .22 for self-report (n = 13) versus .28 for semistructured interview (n = 4).

The overall median values were also disaggregated based on specific, systematic methods (e.g., the LEAD method) or specific instruments (e.g., SWAP) that were employed to collect clinician ratings, in order to determine if this moderated agreement. The bottom of Table 2 presents the median agreement for specific PD diagnoses from studies where the clinical treatment team had collaboratively diagnosed PDs using the LEAD method. The median kappa values for individual PDs ranged from .30 to .65 across the studies, with an overall median of .33. This value was slightly higher than the overall median across all other studies (κ = .26), suggesting that the use of the LEAD method did increase cross-method convergence.

Similarly, this review sought to determine whether the use of a version of the SWAP affected the convergence with other sources. The median convergent correlation across the five studies that employed the SWAP ranged from .24 to .35, with an overall median of .33. This median dimensional agreement for a specific PD did appear somewhat higher than the overall median, consistent with the increase for the LEAD method. Finally, the last row presents the overall median values from all other studies (i.e., those that did not use the SWAP or the LEAD

method). This does suggest that using these more systematic methods modestly increased the level of agreement between clinician ratings and other sources.

Discussion

An important finding from the nine published studies that have reported on the agreement between PD diagnostic ratings assigned two independent clinicians is that interrater reliability is modest across a variety of methods ($\kappa = .40$; $r = .46$). Perhaps surprisingly, it seems to make little difference whether clients are interviewed jointly or separately as both estimates were quite similar. The overall median kappa for agreement between individual PD diagnoses was .40. Although some would likely debate the thresholds they assigned, this level of interrater reliability would be considered acceptable by Kraemer et al. (2012).

To provide context for this value it is instructive to compare it to agreement estimates from other fields. For example, in consulting Meyer and colleagues (2001) review, this was comparable to the meta-analytic agreement observed between traditional dental X-rays and the diagnosis of between-tooth cavities (Vanrijkom & Verdonschot, 1995). Nonetheless, it is also important to note that the estimated interrater reliability in the present review may also be an overestimate of typical practice, because many studies only reported agreement statistics for those PDs that had sufficient base rates. Thus, one might expect that in a typical practice setting where base rates are unknown, that the agreement between two clinicians describing the same patient might be even lower.

An additional note on the interrater reliability is that it might increase when clinician ratings were recorded in a more systematic and detailed way, but more research is needed. The median value for the two studies that used the SWAP was .61 as compared to .43 for all other methods. However, even this median value might not be representative as there were substantial

differences across the two studies that used the SWAP. The first, Westen and Muderrisoglu (2003), obtained SWAP-200 ratings provided by treating clinicians from local clinics affiliated with the research group. Each of the 16 client participants was rated by their treating clinician and then by the researchers following a 3-hour unstructured clinical interview. The agreement for the PD ratings within that sample was remarkably high, with correlations ranging from .55 (paranoid) to .86 (antisocial), with a median of .74. A more recent study that employed the SWAP in a similar manner within a sample of 145 outpatients reported values that were substantially lower (Westen, Shedler, Bradley, & DeFife, 2012). In that study, interrater reliability correlations for individual PDs ranged from .45 to .59, with a median of .48. Thus, it appears that although use of a SWAP instrument improves interrater reliability, the gains are likely not as dramatic as suggested by the preliminary findings of Westen and Muderrisoglu (2003). Additional research that clarifies this inconsistency across studies using the SWAP would be informative.

In any event, it is reasonable to hypothesize that the increased time spent completing a standardized, lengthier, and more psychometrically sound measures will pay dividends in terms of agreement across clinicians. More research is clearly needed to test this hypothesis and to with other systematic measures beyond the SWAP. For example, research that has two treating clinicians describe the same patient using an informant version of the DSM-5 Section III dimensional trait model (e.g., Markon, Quilty, Bagby, & Krueger, 2013) would be instructive in this regard.

Agreement of Clinicians with other Sources

The present review provides the best available estimate regarding the degree of agreement between treating clinicians' PD diagnoses and those from alternative sources (e.g.,

self-report and interview). The overall median agreement for individual PDs across a variety of studies was κ = .26 (categorical) and *r* = .23 (dimensional). As with interrater agreement, it did appear that the more systematic assessments such as LEAD diagnosis or a version of the SWAP, improved the rate of cross-method agreement, but only modestly. Further, the level of cross-method agreement was largely similar across criterion sources, with slightly better agreement between clinicians' diagnoses and researcher conducted semistructured interviews, than for self-report questionnaires completed by clients. The finding of greater overlap between clinicians and semistructured interviews is perhaps not particularly surprising as they do share some method variance (i.e., clinical judgment). Nonetheless, considering both semistructured interviews and unstructured diagnoses are variants of clinician ratings, perhaps the bigger surprise is that the overall agreement between those two methods was still so modest (mdn *r* = .28).

This overall level of agreement between clinicians and other sources (*r* = .23) has far-reaching implications for the translation of research concerning empirically-based treatments into clinical practice settings. Routine discordance between clinician-generated PD diagnoses and those from research methods raises important questions about whether practicing clinicians can have confidence that the diagnoses they generate will inform treatment selection and planning. For instance, although research has identified effective treatments for specific PDs, the patient groups in those treatment studies are routinely diagnosed using semistructured diagnostic interviews. Thus, the applicability of the treatment outcome research to clinical practice might be questioned. In other words, a clinician should have little confidence that the individual client they diagnose with borderline PD, for example, will necessarily benefit from dialectical behavior therapy despite its considerable empirical support for treating the condition (Linehan, Tutek, Heard, & Armstrong, 1994).

It is also worth noting that the overall median dimensional agreement reported here (.23) approximates, but is somewhat lower than the median correlation of .36 reported between informants and self-report measures of PDs by Klonsky and colleagues (2002). Considering that clinicians are a specific type of informant, one can begin to make comparisons to the agreement observed for other types of informants. Interestingly, this would suggest that clinicians' formulations of their clients' personalities share less common variance with the client's own self-description than do reports by spouses, peers, or other informants. One potential explanation for this finding would be that clinicians spend only an hour per week with the client, in one specific context, whereas spouses spend considerably more time with the client in a vast array of contexts. More work that contrasts the incremental validity of clinician and spouse informants would be quite valuable. In any event, the present results indicate that even in the most optimistic scenario—when clinician ratings collected via systematic method are compared to a semistructured interview—the shared variance between clinicians' diagnoses and other methods would amount to only 10-12%.

Comparative Validity

Considering the limited amount of overlapping variance between clinicians and other sources, a fundamentally important question concerns the relative validity of the sources. In other words, when faced with conflicting information from a self-report questionnaire and a clinicians' diagnosis which source is more accurate or valid for predicting important external criteria. Some have argued for the fundamental validity of clinicians' PD ratings (Westen, 1997) due to their extensive training and refined clinical judgment as well as concerns about the validity of client-reported PD ratings (Huprich, Bornstein, & Schmitt, 2011; Westen, 1997) [although for an alternative viewpoint see Rogers (2003), Widiger and Boyd (2009), or

Zimmerman (2003)]. For example, Huprich and Bornstein (2007) argued that self-report questionnaires completed by clients have significant limitations and might be inaccurate due to PD clients' lack of insight into problematic aspects of their personalities or even deliberate attempts to portray themselves in positive or negative ways. Thus, one might reasonably hypothesize that clinicians would provide diagnostic ratings that cumulatively contained more valid variance for predicting important outcomes such as ones interpersonal or occupational functioning.

Then again, research concerning the relative validity of PD ratings by self and informants would suggest that there may actually be reciprocal validity such that each source increments the other. Klein (2003) compared the PD ratings from clients and a knowledgeable informant within a group of 85 depressed outpatients. He reported that when the clients were reassessed 7.5 years later, both sources incrementally predicted depression scores and global functioning assessed by a neutral method (i.e., semistructured interviews) above the other. Thus, he concluded that "both patients' and informants' reports of personality disorders and personality disorder features make independent and unique contributions to predicting symptomatology, social adjustment, and global functioning" (p. 221). Subsequent research in cross-sectional samples has provided similar results (Miller, Pilkonis, & Clifton, 2005). Thus, one might well expect that clinicians and other methods would also increment one another. Perhaps surprisingly, the existing research does not support such reciprocity.

A recent report from the CLPS does shed light upon the question of the relative validity of clinicians' PD diagnoses and those provided by self-report questionnaires and semistructured interviews. Samuel et al. (2013) examined 320 participants with available baseline PD ratings provided by their treating clinician via the PAF (Shea et al., 1987), as well as a semistructured

interview (Diagnostic Interview for DSM-IV Personality Disorders; Zanarini, Frankenburg, Sickel, & Yong, 1996), and a self-report questionnaire (i.e., PD scales from the Schedule for Nonadaptive and Adaptive Personality; Clark, Simms, Wu, & Casillas, in press). In order to avoid content overlap and methodological confounds they examined the ability of these three sets of diagnostic ratings to predict two assessments of psychosocial functioning (one assessed via self-report and one a semistructured interview) from the 5-year follow-up. For example, the clinician ratings from the PAF and the SNAP-2 PD scales were entered in a hierarchical regression model to predict psychosocial functioning as assessed by a semistructured interview. In this way, the criteria and outcome were neutral with respect to method.

The results of these analyses indicated that PD ratings from both the semistructured interview and the self-report questionnaire routinely incremented the prediction offered by the clinician ratings, but the reverse was only rarely the case. In subsequent analyses with a subset of the clinicians who were extremely familiar with the patients (i.e., at least 1 year of treatment) produced a similar pattern of results. This provocatively suggested that self-report questionnaires and semistructured interviews should be seen as more valuable than ratings by treating clinicians for the prospective prediction of psychosocial functioning. Although the findings from Samuel and colleagues (2013) beg for additional research that uses other instruments, other samples, and other independent criteria, they suggest that, at minimum, the disagreement between clinicians' ratings and other methods are not purely a function of limitations associated with self-report.

In this regard, a broader literature is emerging that suggests the limitations of self-report for personality pathology may be less of a concern than previously supposed. Recent studies have suggested that ratings for narcissism and psychopathy, two of the PDs most often implicated as potentially biased due to lack of insight or faking good, have supported the validity

of self-report assessments. A meta-analysis of a variety of psychopathy scales indicates that self-report results are not strongly related to indicators of positive response bias (Ray et al., 2013). Further, studies have indicated that individuals are able to detect their own levels of narcissism and psychopathy in ways that agree with external ratings (Carlson, Vazire, & Oltmanns, 2011; Miller, Jones, & Lynam, 2011). In short, these findings counter suggestions about the invalidity of self-report questionnaires or semistructured interviews for PD diagnosis (Huprich et al., 2011) and even suggest their increased use in clinical practice.

Explanations

In some respects, the limited agreement between clinicians' PD ratings and other methods may not be particularly surprising, as there are important method differences across these sources. Whereas questionnaires and structured interviews are designed to be comprehensive in terms of their coverage of personality pathology and systematic in their assessment of individual criteria, clinicians' PD diagnoses are typically informed by unstructured interviews that are likely idiosyncratic (Zimmerman & Mattia, 1999) and neglect the full list of diagnostic criteria (Westen & Weinberger, 2004).

Thus, although it is certainly possible that clients' self-reports are limited, there are also compelling reasons to believe clinicians reports are equally, if not more limited. An extensive literature has suggested that clinicians are imperfect at collecting and organizing information obtained during clinical interviews (Grove, Zald, Lebow, Snitz, & Nelson, 2000; Meehl, 1954). Clinicians' ratings also might be limited practically by the fact that they rely on observed behaviors to make diagnoses, yet typically interact with their patients in only a single setting (i.e., the consulting room) that has proscribed social roles and restricts patients' behavioral repertoires. However, even if clinicians obtain all relevant information during an interview,

cognitive biases may enter during transcription and encoding. For example, research has found that salient features are more heavily weighted than others (Blashfield & Herkov, 1996; Morey & Ochoa, 1989). In other words, clinicians appear likely to assign the diagnosis of borderline PD to a client who presents with affective lability and self-injurious behavior without conducting a full assessment of the other seven criteria. It is understandable that such tendencies might also contribute to errors in PD diagnoses including the well-documented tendency toward biases according to gender (Anderson, Sankis, & Widiger, 2001; Flanagan & Blashfield, 2003; Samuel & Widiger, 2009).

In addition to these methodological and procedural differences, clinicians' PD ratings are often aggregated informally whereas questionnaires are scored according to predefined algorithms (Westen & Weinberger, 2004). The results of the present review did suggest that when clinicians utilize formal rating scales such as the SWAP-200, the interrater reliability and the convergence with other sources increased modestly (i.e., correlations increased by approximately .10). Although this increase certainly supports clinicians' use of systematic assessments when providing PD diagnoses, it does not appear that this change alone is sufficient to fundamentally alter existing findings. It is also important to note, in this regard, that this review considered only the *convergent* validity of clinicians' ratings with other sources. It is quite possible, for example, that the use of a more psychometrically robust measure simply increases the shared variance with a variety of criteria and also would elevate discriminant correlations. This is one possible explanation for the elevated interrater correlations presented by Westen and Muderrisoglu (2003). Indeed, the authors of the SWAP have subsequently eliminated the q-sort fixed-distribution (Blagov, Bi, Shedler, & Westen, 2012) to avoid concerns that it might artificially inflate convergent and discriminant validity. Thus, future studies that

compare the convergent and discriminant validity of more systematic clinician ratings with other methods would be quite informative in isolating this possibility.

Integrating Clinicians' Ratings with those from Other Sources

More generally, it is unlikely that any one source of personality description should be preferred globally over another (Connelly & Ones, 2010). Instead, a more fruitful research question is "who can most validly tell us what about whom, under which conditions?" Connelly and Ones (2010) meta-analyzed a wide literature on the accuracy of observer ratings of personality and found that although they were quite useful for predicting behavior, relative to self-report, there were important nuances. Interestingly, it was the level of interpersonal intimacy between the observer and the target, not the frequency or duration of contact that increased the accuracy of self and other ratings. Extrapolated into a diagnostic context this suggests that although simply seeing a client for a greater number of sessions or even longer sessions (i.e., familiarity) may not increase the therapist's ability to provide accurate personality ratings, the cross-method convergence might be increased by a stronger rapport or working alliance.

Moreover, Connelly and Ones (2010) also found that there were important differences across aspects of personality, such that informants had greater accuracy for the more observable traits, such as extraversion and conscientiousness. In contrast, more internal aspects of personality such as negative emotionality and openness were more difficult for observers to assess accurately. Interestingly, Ready, Clark, Watson, and Westerhouse (2000) suggested that when rating more difficult-to-judge traits, informants tend to display a "self-based heuristic" such that they rate the target more similarly to their own self-description. Vazire's (2010) work on Self-Other Knowledge Asymmetry (SOKA) also suggests that the observability of a construct is important for understanding self and other agreement, but also highlights the importance of

evaluativeness. Specifically, those traits or features that are more evaluative in nature (e.g., attractiveness, intelligence) may moderate agreement such that well-acquainted others may be more accurate than the self-perception. This is highly relevant to PDs as many pathological traits would be said to be somewhat evaluative and suggests a future line of inquiry.

In this way, future studies should explore the possibility that clinicians and clients can each provide valid (and perhaps complementary) information about different aspects of personality pathology. For example, it may well be that clinicians are well-equipped to provide ratings for more observable aspects of personality, such as impulsivity, but are relatively less accurate at detecting internal processes, such as dissociation. Such findings would dovetail with available data for self-report and semistructured interview assessments in the PD field (Hopwood et al., 2008). Hopwood and colleagues (2008) noted that among measures of borderline PD a self-report questionnaire exhibited greater accuracy than a semistructured interview for criteria that required access to internal states (e.g., chronic emptiness), whereas the reverse was true for more explicitly observable indicators (i.e., self-harm). In this way, research that begins to answer how the judgment and objectivity of a clinician, as well as the client's expertise on him or herself, can be harnessed most effectively to provide the most valid picture of a client would be extremely useful. In sum, additional research is needed to determine who can provide the most accurate information about what aspects of a client's personality. Such data would be quite valuable for informing clinical diagnosis and care.

Is Diagnostic Disagreement Specific to PDs?

Finally, it should be noted that the modest agreement between clinicians' diagnoses and those from other methods is not specific to PDs. In fact, similar rates of cross-method agreement have been reported for a variety of other psychiatric diagnoses (Rettew, Lynch, Achenbach,

Dumenci, & Ivanova, 2009). Rettew and colleagues (2009) aggregated 38 studies reporting the agreement between clinician-generated diagnoses and those from structured diagnostic interviews (including three on PDs that were included in the present review). Rettew and colleagues (2009) reported categorical agreement of .27 across the diagnostic manual. There were specific disorders and diagnostic categories that appeared to fare better (e.g., $\kappa = .84$ for anorexia nervosa) and others worse (e.g., $\kappa = .14$ for affective disorders), but the overall value was not all that dissimilar than what we report here for PDs. Thus, it should not be implied that diagnostic discordance between practicing clinicians and other sources is peculiar to PDs. This raises the concern that diagnostic discrepancy between research and clinical settings may limit the opportunity for evidence-based practice across all, or most, of psychopathology.

Moving toward a Dimensional Model

It is important to note that this review focused on clinicians' diagnoses of the traditional PD categories. Although these PD categories were retained, verbatim, as the official nomenclature within DSM-5, a number of compelling concerns have been raised about the categorical model (e.g., Clark, 2007; Trull & Durrett, 2005; Widiger & Samuel, 2005a). Some of these concerns fundamentally affect their validity, such as excessive heterogeneity and problematic co-occurrence. Thus, it is also possible that the level of agreement observed here is at least partially suppressed by the unreliability and invalidity of the categories themselves.

It is worth considering, in this regard, that DSM-5 also adopted an alternative model of PDs within Section III to encourage further research. Specifically, the section III hybrid system includes a dimensional trait model that has five higher-order domains that are consistent with the five-factor model (Gore & Widiger, 2013; Thomas et al., 2013) and other trait models (Harkness, Finn, McNulty, & Shields, 2012). There have been five studies that examined clinicians'

application of dimensional trait models to their clients and can clarify the agreement across methods/sources.

Soldz, Budman, Demby, and Merry (1995) obtained self-report and clinician descriptions of the five-factor model (FFM) for 35 outpatients and reported that the level of agreement for the five domains ranged from .22 (neuroticism) to .56 (extraversion), with an overall median of .13. Piedmont and Ciarrocchi (1999) reported somewhat higher agreement between 132 outpatients and their treating therapists, with convergent correlations ranging from .23 (conscientiousness) to .46 (openness), with a median of .28. Samuel and Widiger (2010) is the only study that has considered multiple sources simultaneously as they compared clinician ratings of the FFM to scores from a semistructured interview (SIFFM; Trull & Widiger, 1997) as well as the self- and informant report versions of the NEO Personality Inventory – Revised (Costa & McCrae, 1992). Samuel and Widiger noted differences in agreement across these sources as the median agreement with the treating therapists' ratings were .35 for the semistructured interview, .20 for self-report, and .11 for informant-report. Finally, Few et al. (2010) utilized the LEAD method to provide FFM ratings and then compared these to independent self-report ratings of trait dimensions from the SNAP-2 (Clark et al., in press). They reported that agreement was .45 between negative temperament and neuroticism; -.32 between disinhibition and conscientiousness (negative correlations indicate convergence between these two traits); and .52 between positive temperament and extraversion. A more recent study by Few and colleagues (2013), did not collect ratings by treating clinicians but is still relevant. Few and colleagues explored the agreement between the DSM-5 Section III dimensional traits as rated by research clinicians following an interview and those provided by the client on the PID-5. The overall

agreement for the five domains ranged from .50 (psychoticism) to .68 (negative affectivity), with a median of .63.

Taken together, this emerging literature on the clinical use of dimensional trait models appears to mirror the larger literature on the PD categories, but with perhaps slightly higher convergence overall. This suggests that a) weak convergence between clinicians and self-report is not attributable solely to the model being used and b) more structured, systematic methods still produce higher agreement. Additional research that investigates the agreement between therapists and clients using the FFM that appears in DSM-5 Section III is sorely needed to better answer this question, but initial results with dimensional trait models appear to hold some promise of improved convergence.

Methodological Limitations of the Literature

The current review summarizes a wide and disparate literature on the agreement between PD diagnoses provided by treating clinicians and alternative diagnostic methods. Although prior reviews have covered portions of this literature, this effort was focused on externally valid diagnostic practices and is comprehensive in its coverage of relevant studies to provide the best available estimate of this agreement. Nonetheless, there are a number of factors that limit the contribution. First, the disparate nature of this literature precluded a formal aggregation of weighted effect sizes that would account for the sample size and the quality of available data points. Nonetheless, this is the approach that was used by Klonsky et al (2002) to summarize the agreement between self and informant descriptions of PDs and it has become a seminal source on that topic. Nonetheless, it would be ideal if future studies were more systematic in their reporting of standard agreement statistics.

It is also important to note that due to limitations of the extant literature it was not possible to summarize agreement statistics for individual PD diagnoses. Instead, the median values of all PDs examined in given studies were used to create the summary variable. In this regard it is quite possible that there are important variations across the *DSM-IV* PD constructs in terms of their agreement across sources, such that agreement is substantially higher for some PDs than others. At the very least, we can be fairly certain that the values presented here are relatively robust indicators of the general level of agreement across this class of disorders, but future work that isolates specific PDs, traits, or criteria would allow a more precise estimate of potential moderators of this agreement.

Finally, it is crucial to acknowledge that this review considered only convergent validity across sources, whereas a complete context for these results would also require the discriminant validity coefficients. In this regard, it is quite possible that the somewhat higher agreement observed for more structured methods of collecting therapist ratings, such as the SWAP-200, might simply reflect a greater psychometric quality that produces higher correlations with target and non-target constructs. Future research that considers the ratio of convergent to discriminant correlations would be most informative in regard to the ultimately diagnostic utility of any given method or source.

Conclusions and Practical Implications

A primary conclusion from the literature reviewed here is that PD diagnoses assigned by treating clinicians evince moderate interrater reliability and relate modestly with other sources typically used in research settings. On the one hand, this finding portends poorly for the validity of PD diagnoses assigned within routine clinical practice and is likely to hold back efforts to translate empirical treatment findings to practice settings. On the other, it does not appear that

these findings are completely unique to PDs, as interrater reliability and cross-method convergence appear problematic across the diagnostic manual. Nonetheless, a more complete understanding of the relative understanding of the predictive validity of PD diagnoses across sources, controlling for method, is necessary to inform diagnostic practices. The inclusion of an alternative, dimensional model of PDs in DSM-5 holds the promise of improving the conceptualization of personality pathology by building upon the foundation of basic science. Yet, emerging findings suggest that this change alone might not fully remediate clinicians' diagnostic agreement with other sources. Future research that determines how clinicians can most accurately capture information relevant to forming valid PD diagnoses would be highly informative. In the interim though, a practical suggestion is the routine use of self-report questionnaires as a first step toward diagnosis. Self-report questionnaires are freely available to practitioners, are relatively easy to implement, and prior concerns about invalid responding has been alleviated. Thus, a standard inclusion of a PD assessment tool, such as the Personality Inventory for DSM-5 (Krueger, Derringer, Markon, Watson, & Skodol, 2012), early in treatment appears likely to be beneficial for improving the valid identification of personality pathology within clinical practice. Ultimately, research which explores how alternative sources can be combined to produce the most valid assessment of personality pathology is crucial for bridging the promise of that science to everyday clinical practice.

References

Achenbach, T. M., Krukowski, R. A., Dumenci, L., & Ivanova, M. Y. (2005). Assessment of Adult Psychopathology: Meta-Analyses and Implications of Cross-Informant Correlations. *Psychological Bulletin, 131*(3), 361-382. doi: 10.1037/0033-2909.131.3.361

Allard, K., & Grann, M. (2000). Personality disorders and patient-informant concordance on DIP-Q self-report in a forensic psychiatric inpatient setting. *Nordic Journal of Psychiatry, 54*(3), 195-200. doi: Doi 10.1080/080394800750019105

Anderson, K. G., Sankis, L. M., & Widiger, T. A. (2001). Pathology versus statistical infrequency: Potential sources of gender bias in personality disorder criteria. *Journal of Nervous and Mental Disease, 189*(10), 661-668. doi: Doi 10.1097/00005053-200110000-00002

Andreas, S., Theisen, P., Mestel, R., Koch, U., & Schulz, H. (2009). Validity of routine clinical DSM-IV diagnoses (AXIS I/II) in inpatients with mental disorders. *Psychiatry Research, 170*(2-3), 252-255. doi: 10.1016/j.psychres.2008.09.009

Bender, D. S., Dolan, R. T., Skodol, A. E., Sanislow, C. A., Dyck, I. R., McGlashan, T. H., . . . Gunderson, J. G. (2001). Treatment utilization by patients with personality disorders. *American Journal of Psychiatry, 158*(2), 295-302. doi: DOI 10.1176/appi.ajp.158.2.295

Bieling, P. J., MacQueen, G. M., Marriot, M. J., Robb, J. C., Begin, H., Joffe, R. T., & Young, L. T. (2003). Longitudinal outcome in patients with bipolar disorder assessed by life-charting is influenced by DSM-IV personality disorder symptoms. *Bipolar Disorders, 5*(1), 14-21.

Blagov, P. S., Bi, W., Shedler, J., & Westen, D. (2012). The Shedler-Westen Assessment Procedure (SWAP): Evaluating Psychometric Questions About Its Reliability, Validity,

and Impact of Its Fixed Score Distribution. *Assessment, 19*(3), 370-382. doi: Doi 10.1177/1073191112436667

Blashfield, R. K., & Herkov, M. J. (1996). Investigating clinician adherence to diagnosis by criteria: A replication of Morey and Ochoa (1989). *Journal of Personality Disorders, 10*(3), 219-228.

Bradley, R., Hilsenroth, M., Guarnaccia, C., & Westen, D. (2007). Relationship between clinician assessment and self-assessment of personality disorders using the SWAP-200 and PAI. *Psychological Assessment, 19*(2), 225-229. doi: 10.1037/1040-3590.19.2.225

Bronisch, T., Flett, S., Garciaborreguero, D., & Wolf, R. (1993). Comparison of a self-rating questionnaire with a diagnostic checklist for the assessment of DSM-III-R personality disorders. *Psychopathology, 26*(2), 102-107.

Bronisch, T., Garciaborreguero, D., Flett, S., Wolf, R., & Hiller, W. (1992). The Munich Diagnostic Checklist for the Assessment of DSM III-R Personality Disorders for use in routine clinical care and research. *European Archives of Psychiatry and Clinical Neuroscience, 242*(2-3), 77-81. doi: Doi 10.1007/Bf02191550

Cantrell, J. D., & Dana, R. H. (1987). Use of the Millon Clinical Multiaxial Inventory (Mcmi) as a Screening Instrument at a Community Mental-Health-Center. *Journal of Clinical Psychology, 43*(4), 366-375. doi: Doi 10.1002/1097-4679(198707)43:4<366::Aid-Jclp2270430405>3.0.Co;2-P

Carlson, E. N., Vazire, S., & Oltmanns, T. F. (2011). You Probably Think This Paper's About You: Narcissists' Perceptions of Their Personality and Reputation. *Journal of Personality and Social Psychology, 101*(1), 185-201. doi: Doi 10.1037/A0023781

Chick, D., Sheaffer, C. I., Goggin, W. C., & Sison, G. F. (1993). The relationship between MCMI personality scales and clinician-generated DSM-III—R personality disorder diagnoses. *Journal of Personality Assessment, 61*(2), 264-276. doi: 10.1207/s15327752jpa6102_8

Clark, L. A. (2007). Assessment and diagnosis of personality disorder: Perennial issues and an emerging reconceptualization. *Annual Review of Psychology, 58,* 227-257. doi: DOI 10.1146/annurev.psych.57.102904.190200

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.

Connelly, B. S., & Ones, D. S. (2010). An other perspective on personality: Meta-Analytic integration of observers' accuracy and predictive validity. *Psychological Bulletin, 136*(6), 1092-1122. doi: Doi 10.1037/A0021212

Costa, P. T., Jr., & McCrae, R. R. (1992). *Professional Manual: Revised NEO Personaltiy Inventory and NEO Five-Factor Inventory*. Odessa, FL: PAR, Inc.

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

Dreessen, L., & Arntz, A. (1999). Personality disorders have no excessively negative impact on therapist-rated therapy process in the cognitive and behavioural treatment of Axis I anxiety disorders. *Clinical Psychology & Psychotherapy, 6*(5), 384-394. doi: 10.1002/(sici)1099-0879(199911)6:5<384::aid-cpp218>3.0.co;2-8

Eaton, W. W., Martins, S. S., Nestadt, G., Bienvenu, O. J., Clarke, D., & Alexandre, P. (2008). The Burden of Mental Disorders. *Epidemiologic Reviews, 30*(1), 1-14. doi: DOI 10.1093/epirev/mxn011

Egan, S., Nathan, P., & Lumley, M. (2003). Diagnostic concordance of ICD-10 personality and comorbid disorders: a comparison of standard clinical assessment and structured interviews in a clinical setting. *Australian and New Zealand Journal of Psychiatry, 37*(4), 484-491. doi: DOI 10.1046/j.1440-1614.2003.01226.x

Few, L. R., Miller, J. D., Morse, J. Q., Yaggi, K. E., Reynolds, S. K., & Pilkonis, P. A. (2010). Examining the reliability and validity of clinician ratings on the Five-Factor Model Score Sheet. *Assessment, 17*(4), 440-453. doi: 10.1177/1073191110372210

Few, L. R., Miller, J. D., Rothbaum, A. O., Meller, S., Maples, J., Terry, D. P., . . . MacKillop, J. (2013). Examination of the Section III DSM-5 Diagnostic System for Personality Disorders in an Outpatient Clinical Sample. *Journal of Abnormal Psychology, 122*(4), 1057-1069. doi: Doi 10.1037/A0034878

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

Fountoulakis, K. N., Leucht, S., & Kaprinis, G. S. (2008). Personality disorders and violence. *Current Opinion in Psychiatry, 21*(1), 84-92.

Fridell, M., & Hesse, M. (2006). Clinical diagnosis and SCID-II assessment of DSM-III-R personality disorders. *European Journal of Psychological Assessment, 22*(2), 104-108. doi: Doi 10.1027/1015-5759.22.2.104

Gore, W. L., & Widiger, T. A. (2013). The DSM-5 dimensional trait model and five-factor models of general personality. *Journal of Abnormal Psychology, 122*(3), 816-821.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12*(1), 19-30. doi: 10.1037/1040-3590.12.1.19

Gunderson, J. G., Stout, R. L., McGlashan, T. H., Shea, T., Morey, L. C., Grilo, C. M., . . . Skodol, A. E. (2011). Ten-year course of borderline personality disorder psychopathology and function from the Collaborative Longitudinal Personality Disorders Study. *Archives of General Psychiatry*, *68*(8), 827-837. doi: DOI 10.1001/archgenpsychiatry.2011.37

Harkness, A. R., Finn, J. A., McNulty, J. L., & Shields, S. M. (2012). The Personality Psychopathology-Five (PSY-5): recent constructive replication and assessment literature review. *Psychological Assessment, 24*(2), 432-443. doi: 10.1037/a0025830

Hesse, M. (2005). Social workers' ratings of comorbid personality disorders in substance abusers. *Addictive Behaviors, 30*(6), 1241-1246. doi: 10.1016/j.addbeh.2004.12.002

Hesse, M., & Thylstrup, B. (2008). Inter-rater agreement of comorbid DSM-IV personality disorders in substance abusers. *BMC Psychiatry, 8*.

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*(1), 81-85. doi: 10.1037/1040-3590.20.1.81

Huprich, S. K., & Bornstein, R. F. (2007). An overview of issues related to categorical and dimensional models of personality disorders assessment. *Journal of Personality Assessment, 89*(1), 3-15.

Huprich, S. K., Bornstein, R. F., & Schmitt, T. A. (2011). Self-Report methodology is insufficient for improving the assessment and classification of axis II personality disorders. *Journal of Personality Disorders, 25*(5), 557-570. doi: 10.1521/pedi.2011.25.5.557

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

Klein, M. H., Benjamin, L. S., Rosenfeld, R., Treece, C., Husted, J., & Greist, J. H. (1993). The Wisconsin Personality Disorders Inventory: Development, reliability, and validity. *Journal of Personality Disorders, 7*(4), 285-303.

Klonsky, E. D., Oltmanns, T. F., & Turkheimer, E. (2002). Informant-reports of personality disorder: Relation to self-reports and future research directions. *Clinical Psychology-Science and Practice, 9*(3), 300-311.

Kraemer, H. C., Kupfer, D. J., Clarke, D. E., Narrow, W. E., & Regier, D. A. (2012). DSM-5: How reliable is reliable enough? *American Journal of Psychiatry, 169*(1), 13-15. doi: DOI 10.1176/appi.ajp.2011.11010050

Krueger, R. F., Derringer, J., Markon, K. E., Watson, D., & Skodol, A. E. (2012). Initial construction of a maladaptive personality trait model and inventory for DSM-5. *Psychological Medicine, 42*(9), 1879-1890. doi: 10.1017/S0033291711002674

Linehan, M. M., Tutek, D. A., Heard, H. L., & Armstrong, H. E. (1994). Interpersonal outcome of cognitive-behavioral treatment for chronically suicidal borderline patients. *American Journal of Psychiatry, 151*(12), 1771-1776.

Löffler-Stastka, H., Ponocny-Seliger, E., Fischer-Kern, M., Rössler-Schülein, H., Leithner-Dziubas, K., & Schuster, P. (2006). Validation of the SWAP-200 for Diagnosing Psychostructural Organization in Personality Disorders. *Psychopathology, 40*(1), 35-46. doi: 10.1159/000096388

Markon, K. E., Quilty, L. C., Bagby, R. M., & Krueger, R. F. (2013). The development and psychometric properties of an informant-report form of the Personality Inventory for DSM-5 (PID-5). *Assessment, 20*(3), 370-383.

McDermut, W., & Zimmerman, M. (2005). Assessment instruments and standardized evaluation. In J. Oldham, A. E. Skodol & D. Bender (Eds.), *The American Psychiatric Publishing Textbook of Personality Disorder*. Washington, DC: American Psychiatric Publishing.

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

Meier, P., & Barrowclough, C. (2009). Mental health problems: Are they or are they not a risk factor for dropout from drug treatment? A systematic review of the evidence. *Drugs-Education Prevention and Policy, 16*(1), 7-38. doi: Pii 908612145

Doi 10.1080/09687630701741030

Mellsop, G., Varghese, F., Joshua, S., & Hicks, A. (1982). The reliability of Axis II of DSM-III. *American Journal of Psychiatry, 139*(10), 1360-1361.

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., . . . Reed, G. M. (2001). Psychological testing and psychological assessment - A review of evidence and issues. *American Psychologist, 56*(2), 128-165. doi: Doi 10.1037//0003-066x.56.2.128

Miller, J. D., Jones, S. E., & Lynam, D. R. (2011). Psychopathic Traits From the Perspective of Self and Informant Reports: Is There Evidence for a Lack of Insight? *Journal of Abnormal Psychology, 120*(3), 758-764. doi: Doi 10.1037/A0022477

Miller, J. D., Pilkonis, P. A., & Clifton, A. (2005). Self- and other-reports of traits from the five-factor model: Relations to personality disorder. *Journal of Personality Disorders, 19*(4), 400-419. doi: DOI 10.1521/pedi.2005.19.4.400

Molinari, V., Kunik, M. E., Mulsant, B., & Rifai, A. H. (1998). The relationship between patient, informant, social worker, and consensus diagnoses of personality disorder in elderly depressed inpatients. *The American Journal of Geriatric Psychiatry, 6*(2), 136-144.

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

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

Newton-Howes, G., Tyrer, P., & Johnson, T. (2006). Personality disorder and the outcome of depression: meta-analysis of published studies. *British Journal of Psychiatry, 188*, 13-20.

Noren, K., Lindgren, A., Haellstom, T., Thormaehlen, B., Vinnars, B., Wennberg, P., . . . Barber, J. P. (2007). Psychological distress and functional impairment in patients with personality

disorders. *Nordic Journal of Psychiatry*, *61*(4), 260-270. doi: Doi 10.1080/08039480701414973

North, C. S., Pollio, D. E., Thompson, S. J., Ricci, D. A., Smith, E. M., & Spitznagel, E. L. (1997). A comparison of clinical and structured interview diagnoses in a homeless mental health clinic. *Community Mental Health Journal*, *33*(6), 531-543. doi: 10.1023/a:1025052720325

Perry, J. C. (1992). Problems and Considerations in the Valid Assessment of Personality Disorders. *American Journal of Psychiatry*, *149*(12), 1645-1653.

Piedmont, R. L., & Ciarrocchi, J. W. (1999). The utility of the revised NEO personality inventory in an outpatient, drug rehabilitation context. *Psychology of Addictive Behaviors*, *13*(3), 213-226.

Piersma, H. L. (1987). The MCMI as a measure of DSM-III Axis II diagnoses: An empirical comparison. *Journal of Clinical Psychology*, *43*(5), 478-483. doi: 10.1002/1097-4679(198709)43:5<478::aid-jclp2270430508>3.0.co;2-z

Pilkonis, P. A., Heape, C. L., Proietti, J. M., Clark, S. W., Mcdavid, J. D., & Pitts, T. E. (1995). The Reliability and Validity of 2 Structured Diagnostic Interviews for Personality-Disorders. *Archives of General Psychiatry*, *52*(12), 1025-1033.

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

Pinel, P. (1801, 1962). *A treatise on insanity*. New York: Published under the auspices of the Library of the New York Academy of Medicine by Hafner Pub. Co.

Ray, J. V., Hall, J., Rivera-Hudson, N., Poythress, N. G., Lilienfeld, S. O., & Morano, M. (2013). The Relation Between Self-Reported Psychopathic Traits and Distorted Response Styles: A Meta-Analytic Review. *Personality Disorders-Theory Research and Treatment, 4*(1), 1-14. doi: Doi 10.1037/A0026482

Ready, R. E., Clark, L. A., Watson, D., & Westerhouse, K. (2000). Self- and peer-reported personality: Agreement, trait ratability, and the "self-based heuristic". *Journal of Research in Personality, 34*(2), 208-224. doi: DOI 10.1006/jrpe.1999.2280

Regier, D. A., Kaelber, C. T., Roper, M. T., & Rae, D. S. (1994). The ICD-10 clinical field trial for mental and behavioral disorders: Results in Canada and the United States. *Am J Psychiatry, 151*(9), 1340-1350.

Rettew, D. C., Lynch, A. D., Achenbach, T. M., Dumenci, L., & Ivanova, M. Y. (2009). Meta-analyses of agreement between diagnoses made from clinical evaluations and standardized diagnostic interviews. *International Journal of Methods in Psychiatric Research, 18*(3), 169-184. doi: 10.1002/mpr.289

Rogers, R. (2003). Standardizing DSM-IV diagnoses: The clinical applications of structured interviews. *Journal of Personality Assessment, 81*(3), 220-225. doi: Doi 10.1207/S15327752jpa8103_04

Rossi, G., Van den Brande, I., Tobac, A., Sloore, H., & Hauben, C. (2003). Convergent validity of the MCMI-III personality disorder scales and the MMPI-2 scales. *Journal of Personality Disorders, 17*(4), 330-340. doi: 10.1521/pedi.17.4.330.23970

Samuel, D. B., Anez, L. M., Paris, M., & Grilo, C. M. (in press). The convergence of personality disorder diagnoses across different methods among monolingual (Spanish-speaking only)

Hispanic patients in substance abuse treatment. *Personality Disorders: Theory, Research, and Treatment*.

Samuel, D. B., Sanislow, C. A., Hopwood, C. J., Shea, M. T., Skodol, A. E., Morey, L. C., . . . Grilo, C. M. (2013). Convergent and incremental predictive validity of clinician, self-report, and diagnostic interview assessment methods for personality disorders over five years. *Journal of Consulting and Clinical Psychology, 81*(4), 650-659.

Samuel, D. B., & Widiger, T. A. (2009). Comparative gender biases in models of personality disorder. *Personality and Mental Health, 3*(1), 12-25. doi: 10.1002/pmh.61

Samuel, D. B., & Widiger, T. A. (2010). Comparing personality disorder models: Cross-method assessment of the FFM and DSM-IV-TR. *Journal of Personality Disorders, 24*(6), 721-745. doi: 10.1521/pedi.2010.24.6.721

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

Shedler, J., & Westen, D. (1998). Refining the measurement of Axis II: A Q-sort procedure for assessing personality pathology. *Assessment, 5*(4), 333-353. doi: 10.1177/107319119800500403

Skodol, A. E., Rosnick, L., Kellman, D., Oldham, J. M., & Hyler, S. E. (1988). Validating structured DSM-III-R personality disorder assessments with longitudinal data. *American Journal of Psychiatry, 145*(10), 1297-1299.

Smith, S. W., Hilsenroth, M. J., & Bornstein, R. F. (2009). Convergent validity of the SWAP-200 dependency scales. *Journal of Nervous and Mental Disease, 197*(8), 613-618. doi: Doi 10.1097/Nmd.0b013e3181b08d89

Soeteman, D. I., Verheul, R., & Busschbach, J. J. V. (2008). The burden of disease in personality disorders: Diagnosis-specific quality of life. *Journal of Personality Disorders, 22*(3), 259-268.

Soldz, S., Budman, S., Demby, A., & Merry, J. (1995). Personality traits as seen by patients, therapists and other group members: The Big Five in personality disorder groups. *Psychotherapy: Theory, Research, Practice, Training, 32*(4), 678-687. doi: 10.1037/0033-3204.32.4.678

Spitzer, R. L., Forman, J. B. W., & Nee, J. (1979). DSM-III field trials .1. Initial interrater diagnostic reliability. *American Journal of Psychiatry, 136*(6), 815-817.

Tenney, N. H., Schotte, C. K. W., Denys, D. A. J. P., van Megen, H. J. G. M., & Westenberg, H. G. M. (2003). Assessment of DSM-IV personality disorders in obsessive-compulsive disorder: Comparison of clinical diagnosis, self-report questionnaire, and semi-structured interview. *Journal of Personality Disorders, 17*(6), 550-561. doi: DOI 10.1521/pedi.17.6.550.25352

Thomas, K. M., Yalch, M. M., Krueger, R. F., Wright, A. G. C., Markon, K., & Hopwood, C. J. (2013). The convergent structure of DSM-5 personality trait facets and Five-Factor Model trait domains. *Assessment, 20*(3), 308-311.

Torrens, M., Serrano, D., Astals, M., Pérez-Domínquez, G., Sr., & Martín-Santos, R. (2004). Diagnosing comorbid psychiatric disorders in substance abusers: Validity of the Spanish versions of the Psychiatric Research Interview for Substance and Mental Disorders and the Structured Clinical Interview for DSM-IV. *Am J Psychiatry, 161*(7), 1231-1237. doi: 10.1176/appi.ajp.161.7.1231

Trull, T. J., & Durrett, C. A. (2005). Categorical and dimensional models of personality disorder. *Annual Review of Clinical Psychology, 1*, 355-380. doi: 10.1146/annurev.clinpsy.1.102803.144009

Trull, T. J., Jahng, S., Tomko, R. L., Wood, P. K., & Sher, K. J. (2010). Revised Nesarc Personality Disorder Diagnoses: Gender, Prevalence, and Comorbidity with Substance Dependence Disorders. *Journal of Personality Disorders, 24*(4), 412-426.

Trull, T. J., & Widiger, T. A. (1997). *SIFFM: Structured Interview for the Five-Factor Model of Personality, Professional Manual*. Odessa, FL: Psychological Assessment Resources.

Vanrijkom, H. M., & Verdonschot, E. H. (1995). Factors involved in validity measurements of diagnostic tests for approximal caries: A metaanalysis. *Caries Research, 29*(5), 364-370.

Vazire, S. (2010). Who Knows What About a Person? The Self-Other Knowledge Asymmetry (SOKA) Model. *Journal of Personality and Social Psychology, 98*(2), 281-300. doi: Doi 10.1037/A0017908

Vine, R. G., & Steingart, A. B. (1994). Personality disorder in the elderly depressed. *Canadian Journal of Psychiatry-Revue Canadienne De Psychiatrie, 39*(7), 392-398.

Westen, D. (1997). Divergences between clinical and research methods for assessing personality disorders: Implications for research and the evolution of axis II. *American Journal of Psychiatry, 154*(7), 895-903.

Westen, D., & Muderrisoglu, S. (2003). Assessing personality disorders using a systematic clinical interview: Evaluation of an alternate to structured interviews. *Journal of Personality Disorders, 17*(4), 351-369. doi: 10.1521/pedi.17.4.351.23967

Westen, D., & Shedler, J. (1999). Revising and assessing axis II, Part II: Toward an empirically based and clinically useful classification of personality disorders. *American Journal of Psychiatry, 156*(2), 273-285.

Westen, D., Shedler, J., Bradley, B., & DeFife, J. A. (2012). An Empirically Derived Taxonomy for Personality Diagnosis: Bridging Science and Practice in Conceptualizing Personality. *American Journal of Psychiatry, 169*(3), 273-284. doi: DOI 10.1176/appi.ajp.2011.11020274

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

Widiger, T. A., & Boyd, S. (2009). Assessing personality disorders. In J. Butcher (Ed.), *Oxford Handbook of Personality Assessment*. New York: Oxford University Press.

Widiger, T. A., & Samuel, D. B. (2005a). Diagnostic categories or dimensions? A question for the diagnostic and statistical manual of mental disorders-fifth edition. *Journal of Abnormal Psychology, 114*(4), 494-504. doi: 10.1037/0021-843X.114.4.494

Widiger, T. A., & Samuel, D. B. (2005b). Evidence-based assessment of personality disorders. *Psychological Assessment, 17*(3), 278-287. doi: 10.1037/1040-3590.17.3.278

Wilberg, T., Dammen, T., & Friis, S. (2000). Comparing Personality Diagnostic Questionnaire-4+ with longitudinal, expert, all data (LEAD) standard diagnoses in a sample with a high prevalence of axis I and axis II disorders. *Comprehensive Psychiatry, 41*(4), 295-302. doi: DOI 10.1053/comp.2000.0410295

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.

Zimmerman, M. (1994). Diagnosing personality disorders : A review of issues and research methods. *Archives of General Psychiatry, 51*(3), 225-245. doi: 10.1001/archpsyc.1994.03950030061006

Zimmerman, M. (2003). What should the standard of care for psychiatric diagnostic evaluations be? *Journal of Nervous and Mental Disease, 191*(5), 281-286. doi: Doi 10.1097/00005053-200305000-00002

Zimmerman, M., & Mattia, J. I. (1999). Differences between clinical and research practices in diagnosing borderline personality disorder. *American Journal of Psychiatry, 156*(10), 1570-1574.

Table 1

Interrater Reliability of Clinicians Personality Disorder Diagnoses

Study	*n*	PD Definition	Source	Variables Rated	Instrument		Statistic	κ for any PD	Mdn κ	Mdn *r*
Westen et al (2012)	145	SWAP Qs	clinical contact	SWAP items	SWAP-II	separate	*r*			.48
Hesse & Thylstrup (2008)	75	DSM-IV	clinical contact	PDs (dim)	list	separate	*r*			.34
				criteria	checklist	separate	*r*			.46
Hesse (2005)	49	DSM-IV	clinical contact	PDs (dim)	none	separate	*r*			.43
Westen & Muderrisoglu (2003)	16	DSM-IV	clinical contact	SWAP items	SWAP-200	separate	*r*			.74
Molinari et al (1998)	20	DSM-IV	unstructured interview	PDs (dim)	PAF	joint	κ	.44		
			clinical contact	PDs (dim)	PAF	separate	κ	.35		
Regier et al (1994)	491	ICD-10	unstructured interview	PDs (cat)	none	joint	κ	.52	.40	
Bronisch et al	60	DSM-III-	unstructured	individual	MDCL-P	separate	κ	.62	.62	

(1992)		R	interview	criteria						
Mellsop et al (1982)	77	DSM-III	unstructured interview	PDs (cat)	none	separate	κ	.41	.23	
Spitzer et al (1979)	281	DSM-III	unstructured interview	PDs (cat)	none	joint	κ	.61		
						separate	κ	.54		
median kappa								.52	.40	
median correlation										.46

notes: SWAP = Shedler - Westen Assessment Procedure; DSM = Diagnostic and Statistical Manual of Mental Disorders; ICD = International Classification of Diseases; PD = Personality Disorder; PAF = Personality Assessment Form; MDCL-P = Munich Diagnostic Checklist for the assessment of DSM-III-R and ICD-10 Personality Disorders.

Table 2

Convergence of Clinicians Personality Disorder Diagnoses with Other Methods and Sources

Study	*n*	PD Definition	Source	Variables Rated	Clinician Method	Criterion Instrument (Source)	Statistic	κ for any PD	Mdn κ	Mdn *r*
Gritti et al. (in prep)	57	DSM-IV	clinical contact	SWAP items	SWAP-200	MCMI-III (S)	*r*			.35
Samuel et al. (in press)	112	DSM-IV	clinical contact	PDs (dim.)	PAF	PDQ-4 (S)	*r*			.15
	112					DIPD-IV (I)	*r*			.12
Samuel et al. (2013)	320	DSM-IV	clinical contact	PDs (dim.)	PAF	DIPD-IV (I)	*r*		.31	.35
	320					SNAP-2 (S)	*r*		.08	.22
Samuel & Widiger (2010)	77	DSM-IV	clinical contact	PDs (dim.)	DSMRF	PDI-IV (I)	*r*			.24
	86					SNAP (S)	*r*			.08
	61					DSMRF (P)	*r*			.16
Andreas et al. (2009)	55	DSM-IV	intake interview	PDs (cat.)	none	SCID-II (I)	κ		.45	
Smith et al. (2009)	85	DSM-IV	clinical contact	SWAP items	SWAP-200	IIP-64 (S)	*r*			.31
Bradley et al. (2007)	47	DSM-IV	clinical contact	SWAP items	SWAP-200	PAI (S)	*r*			.33
Loffler-Statska et al. (2007)	33	DSM-IV	clinical contact	SWAP items	SWAP-200	SCID-II (I)	*r*			.33

Fridell & Hesse (2006)	138	DSM-III-R	unstructured	criteria	checklist	SCID-II (I)	κ	.48	.26	
Torrens et al. (2004)	105	DSM-IV	clinical contact	PDs (cat.)	LEAD	PRISM (I)	κ		.65	
						SCID-II (I)	κ		.36	
Davidson et al. (2003)	23	SWAP Qs	clinical contact	SWAP items	SWAP-200	SWAP-200 (S)	*r*			.24
Egan, Nathan, and Lumley (2003)	33	ICD-10	clinical contact	PDs (cat.)	none	IPDE (I)	κ	.96	.25	
Rossi et al. (2003)	330	DSM-IV	clinical contact	PDs (dim.)	Rating Form	MCMI-III (S)	*r*			.21
Tenney et al. (2003)	65	DSM-IV	unstructured interview	PDs (cat.)	none	ADP-IV (S)	κ	.15	.36	
						SCID-II (I)	κ	.14	.23	
Allard & Grann (2000)	42	DSM-IV	clinical contact	individual criteria	DIP-Q	DIP-Q (S)	κ	-.07	.03	
							ICC			.07
Wilberg et al. (2000)	100	DSM-IV	clinical contact + SCID-II	PDs (cat.)	LEAD	PDQ-4 (S)	κ		.30	
Dreessen & Arntz (1999)	70	DSM-III-R	clinical contact	PDs (cat.)	TQPP	SCID-II (I)	κ	.12	.03	
North et al. (1997)	97	DSM-III-R	all clinical information	PDs (cat.)	none	DIS (I)	κ		.40	
Chick et al.	101	DSM-III-	clinical contact	individual	checklist	MCMI (S)	*r*			.05

(1993)		R		criteria						
Bronisch et al. (1993)	60	DSM-III-R	unstructured interview	PDs (dim.)	MDCL-P	PDQ-R (S)	κ	.38	.12	
Klein et al. (1993)	103	DSM-III-R	clinical contact	PDs (dim.)	PAF	WISPI (S)	*r*			.36
Vine & Steingart (1994)	64	DSM-III-R	intake interview	PDs (cat.)	none	SCID-II (I)	κ	.38	.23	
Hyler et al. (1989)	552	DSM-III	clinical contact	PDs (dim.)	CAF	PDQ (S)	κ *r*		.08	.31
Morey et al. (1988)	107	DSM-III	discharge chart	PDs (cat.)	none	MMPI-2 (S)	*r*			.19
Skodol et al. (1988)	20	DSM-III-R	clinical contact	PDs (cat.)	LEAD	SCID-II (I)	κ		.30	
Piersma (1987)	43	DSM-III	clinical contact	PDs (cat.)	none	MCMI (S)	κ		.05	
Cantrell & Dana (1987)	72	DSM-III	clinical contact	PDs (cat.)	none	MCMI (S)	κ	.07		
Overall Median								.15	.26	.23
Median Self-								.11	.08	.22

report			
Median			
Interview	.38	.30	.28
LEAD Method	--	.33	--
SWAP Method	--	--	.33
All other			
Methods	.15	.23	.19

Notes: PD = Personality Disorder; Sources are as indicated: S = self-report, I = semistructured interview, and P = peer informant; DSM = Diagnostic and Statistical Manual of Mental Disorders; SWAP = Shedler - Westen Assessment Procedure; cat = categorical; dim = dimensional; MCMI = Millon Clinical Multiaxial Inventory; PAF = Personality Assessment Form; PDQ = Personality Diagnostic Questionnaire; DIPD-IV = Diagnostic Interview for DSM-IV Personality Disorders; SNAP = Schedule for Nonadaptive and Adaptive Personality; PDI-IV = Personality Disorder Interview for DSM-IV; DSMRF = DSM-IV PD Rating Form; SCID-II = Structured Clinical Interview for DSM-IV Axis II; IIP-64 = Inventory of Interpersonal Problems; PAI = Personality Assessment Inventory; LEAD = Longitudinal, Expert, All Data; PRISM = Psychiatric Research Interview for Substance Use and Mental Disorders; ICD = International Classification of Diseases; IPDE = International Personality Disorders Evaluation; ADP-IV = Assessment of DSM-IV Personality Disorder; DIP-Q = DSM-IV and ICD-10 Personality Questionnaire; TQPP = Therapist Questionnaire for Personality Pathology; DIS = Diagnostic Interview Schedule; MDCL-P = Munich Diagnostic Checklist for the Diagnosis of DSM-III-R Personality Disorders; WISPI = Wisconsin Personality Disorders Inventory; CAF = Clinician Assessment Form; MMPI = Minnesota Multiphasic Personality Inventory.