Running Head: Assessing the Assessors

Assessing the Assessors: The Feasibility and Validity of Clinicians as a Source for Personality Disorder Research

Douglas B. Samuel

&

Meredith A. Bucher

In press, Personality Disorders: Theory, Research & Treatment

Authors' Notes:

Douglas B. Samuel and Meredith A. Bucher; Department of Psychological Sciences, Purdue University.

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

The use of knowledgeable informants is a particularly valuable tool for the diagnosis and assessment of personality disorder (PD). This review details the use of one particular type of informant—practicing clinicians—in PD research. We detail a wide variety of studies that have employed clinicians as an assessment source, including those focused on interrater agreement, comparative validity with other methods, cognitive factors of diagnosis, and opinion surveys. We demonstrate limitations, such as potential biases and limited convergent validity, which caution against the assumption that clinicians' ratings should be considered a gold-standard. Nonetheless, we also highlight the potential value of research that focuses on clinicians due to its external validity to real-world practice settings. Finally, we outline several issues to consider when sampling clinicians, such as participation rate and sample size, and call for future research that collects ratings from clinicians using systematic, well-validated measures.

Keywords: *therapist-report; personality pathology; multi-method; diagnosis*

A major methodological decision in any study is choosing the source(s) that will provide ratings of the variables of interest. In a vast majority of cases this rightfully begins with directly asking the individual whose personality and/or psychopathology is being investigated. Nonetheless, the opinions of some other person, an informant, often can provide incrementally useful information. For example, in organizational settings, it is routine to collect information from an employee as well as from the employee's peers and supervisors (Dunning, Heath, & Suls, 2004). Similarly, for personality pathology there are often situations where other sources, such as a knowledgeable informant, can provide information that reveals unique information that increments the self-report (Oltmanns & Turkheimer, 2009). This review covers one particular population of informants—practicing clinicians—and highlights the advantages and potential weaknesses of employing this group in personality disorder (PD) research.

It is important to first highlight the scope of this review. In discussing therapists or clinicians we refer specifically to those trained individuals who provide ratings or judgments on the basis of their clinical interactions. Such distinctiveness in scope and terminology is important as this review is not concerned with diagnoses, ratings, or judgments provided by research personnel, such as LEAD diagnoses assigned by research teams (e.g., Pilkonis, Heape, Ruddy, & Serrao, 1991) or ratings obtained from, or based on, a clinical interview solely for research purposes (e.g., Few et al., 2013). Although these groups are often called "clinicians" or "therapists" in the empirical literature, we believe it is important to reserve these terms for professionals providing clinical care. Similarly, this review is not concerned with studies that employ "expert raters" when the population in question is primarily researchers who provide ratings based on their understanding of the empirical literature. These type of studies are also quite valuable and have been profitably employed for establishing an expert consensus (Mullins-

Sweatt, Bernstein, & Widiger, 2012; Samuel, Lynam, Widiger, & Ball, 2012), aggregating opinions across a broad constituency (Bernstein, Iscan, & Maser, 2007), or providing criteria for construct validation (e.g., Thomas, Wright, Lukowitsky, Donnellan, & Hopwood, 2012). Although the researchers in these studies are often well-trained clinically, professionally licensed, and in many cases are even engaged in the provision of clinical services, their expertise for the ratings comes from their conceptual or empirical contributions to the research literature. The focus of this review is on the opinions and judgments of therapists engaged in clinical practice, as they have unique perspectives.

Advantages of Sampling Clinicians

External Validity. A major advantage of sampling practicing clinicians is the external validity of their ratings. Clinicians are unique in their ability to provide a window into how PDs are diagnosed in clinical practice. In other words, if one seeks to understand the key features and heuristics that drive the cognitive factors of diagnostic decisions (Kim & Ahn, 2002), then one simply must sample practitioners who are actively engaged in this enterprise. It is important to note that this does not necessarily suggest that their diagnoses have construct validity, but at the very least they are reflections of the types of diagnoses that are routinely provided in clinics.

Impartial Third Party. The diagnosis and assessment of personality disorder is particularly challenging in many cases as the disorders are ego-syntonic. PD symptoms might not be considered particularly problematic to the person, but be better reflected by the distress they cause others (Miller, Pilkonis, & Clifton, 2005). Scholars have argued that self-reports of PD pathology are inherently untrustworthy due to concerns about the individual's ability to accurately describe their own personality, either due to lack of insight or deliberate distortion (Ganellen, 2007; Huprich & Bornstein, 2007). To overcome such possible limitations of first-

person accounts, the DSM-5 recommends the use of "supplementary information from other informants" (American Psychiatric Association, 2013, p. 647).

Evidence has repeatedly demonstrated that reports from informants provide incremental validity beyond self-report for predicting concurrent and prospective functioning for a wide variety of outcomes (Connelly & Ones, 2010; Klein, 2003), including military discharge (Fiedler, Oltmanns, & Turkheimer, 2004). In the PD field, significant others and peers are the typical informants (South, Oltmanns, Johnson, & Turkheimer, 2011), although clinicians are one particular type within this broader class.

Although well-acquainted peers are particularly valuable informants as they can provide information about how maladaptive behaviors impact an individual across a variety of situations, their ability to report may be colored by their own perceptions and biases. For example, the spouse of someone with deceitful and manipulative tendencies will likely be quite aware and readily report upon these traits, but might even over-report on them given the likely interpersonal frustrations they have experienced. Clinicians are also humans and countertransference might color their impressions. Nonetheless, their professional positions relative to the client may allow them greater objectivity to accurately portray symptoms.

Training and Clinical Experience. In addition to possible biases, significant others and peers may be limited by their ability to accurately detect the fairly complex patterns of behavior that characterize PDs (Perry, 1992). Although one can reasonably debate the depth of focus on PD diagnosis in most graduate programs, clinicians do receive extensive training on the assessment and diagnosis of mental disorders and can call upon their experience with a wide variety of patients in order to detect and elicit PD-relevant information. Westen (1997) has extensively criticized the abilities of lay persons, including informants, to adequately report on personality

pathology. For example, he noted that "understanding people requires training. Psychiatrists would not need 3- to 5-year residencies if diagnoses of subtle psychological processes were so readily apparent to lay observers" (p. 897). Although it may be tautological to suggest that psychiatrists' impressions of PDs are valid because they completed a psychiatry residency, it is certainly reasonable to suggest that the extensive training, competency, and knowledge that is required to obtain a mental health degree, and ultimate license, does improve one's ability to diagnosis mental illness.

Clinicians Use of PDs in Clinical Practice

In considering the potential value of clinicians as an assessment source, a first step is elucidating how clinicians make PD diagnoses in routine clinical practice. Although DSM-5 (APA, 2013) offers a specific algorithm for arriving at a PD diagnosis—typically meeting half, or one more than half, of the criteria—research suggests clinicians do not apply the DSM-5 so methodically (First et al., 2014). Systematically assessing the diagnostic criteria, such as via a semistructured interview, is the hallmark of PD diagnosis within research settings but is rare in clinical practice settings (Widiger & Samuel, 2005). Instead, Perry (1992) has argued that clinicians prefer to rely on global impressions based on their unstructured interviews and subjective experience of interactions rather than systematically assessing diagnostic criteria. Specifically, clinicians report that they find more value in listening to the patient describe interactions with significant others than in asking direct questions about symptoms or administering formal questionnaires (Westen, 1997). Other studies have argued that these global impressions are themselves based largely on a few diagnostic criteria that are very salient (Kim & Ahn, 2002). But in either case, what is clear is that therapists rarely apply the existing criterion sets systematically when arriving at PD diagnoses (First et al., 2014).

Some have applauded this routine departure from the DSM criterion sets, even going so far as to criticize rigorous adherence to the criteria. Shedler (2015) has argued that the use of criteria "minimized the role of clinical inference and treated PD diagnosis as a largely technical task of tabulating signs and symptoms" (p. 226). Nonetheless, it is the case that when presented with information from a semistructured interview, clinicians readily incorporate this into their diagnoses (Zimmerman & Mattia, 1999). Thus, it appears that clinicians do find value in the systematic application of diagnostic criteria sets.

One criticism of the unstructured diagnostic approach is the possibility, if not probability, that idiosyncratic differences in the aggregation of information limits the reliability and validity of this approach. Westen and Weinberger (2004) have suggested that clinical judgment, per se, should not be conflated with the method of aggregation. They suggest instead that clinicians' diagnoses are simply another source of actuarial prediction when collected using a systematic and comprehensive measurement tool. Westen and Shedler (1999) created the Shedler-Westen Assessment Procedure (SWAP) with such an intention. The SWAP-200 is a clinician-rated card-sort measure that includes 200 statements likely relevant to PD description. These statements are then sorted into a fixed-distribution of eight piles, with successively fewer spaces (Blagov, Bi, Shedler, & Westen, 2012). The item-set uses specialized wording that "provides dynamic clinicians a common vocabulary with which to express their observations and inferences about character and character pathology" (Shedler, 2002; p. 434). As such, the SWAP can only be completed by a treating clinician after a significant period of clinical interaction. The SWAP can be scored for the DSM-5 PD constructs as well as factor-analytically derived prototypes and has obtained validity support including agreement across separate interviewers (Westen & Muderrisoglu, 2003) and relations with outcome variables (Westen, Shedler, Bradley, & DeFife,

2012). Nonetheless, the SWAP has also been criticized on a number of grounds including its fixed distribution (Block, 2008) and other problematic psychometric features (Wood, Garb, Nezworski, & Koren, 2007). Indeed, a recent study has suggested that the increased convergent validity of the SWAP may come at the expense of weakened discriminant validity (Gritti, Samuel, & Lang, in press). More research is needed to investigate the properties of the SWAP-200 as well as other systematic methods of collecting PD descriptions from practicing clinicians.

Construct Validity of Clinicians as an Assessment Source

Interrater Reliability. A central impetus of the specific diagnostic criterion sets that were a primary innovation of DSM-III was enhancing the diagnostic reliability across raters. Kraemer and colleagues (2012) have highlighted how little is known about the rates of diagnostic agreement among independent raters, yet this is central to the validity of any diagnosis. If two separate clinicians interview a client and reach differing conclusions, then one can have little confidence in the overall validity of the diagnoses.

One way to estimate diagnostic agreement of clinician raters is to calculate agreement after they have read a standardized vignette. Such studies have routinely suggested there is broad agreement when clinicians are asked to rate a prototypic case of a PD in terms of the traits of the FFM (Samuel & Widiger, 2004). Similarly, when clinicians are provided with a brief vignette they produce comparable diagnostic ratings (Samuel & Widiger, 2009; Sprock, 2003), providing preliminary support for the interrater reliability of PD diagnoses. Of course, in real-world settings, clinicians are not presented with a brief vignette. Instead, clinicians must meet with a person and elicit diagnostic information on the basis of relatively brief interactions. The DSM-5 field trials evinced a more rigorous methodology, in which two raters separately interviewed a client and the results were mixed with regard to the PD agreement. The rate of pooled agreement

for BPD was considered "good" at *K* = .54, but this combined discrepant values across two sites. The agreement at the Center for Addiction and Mental Health was .75, but the comparable value at the Menninger Clinic was only .34. Moreover, the categorical agreement for antisocial PD -- a construct with historically high reliability -- was only .21. Schizotypal and obsessive-compulsive PDs were not prevalent enough to calculate a reliable estimate (Regier et al., 2013).

Samuel (2015) reviewed the literature and found nine studies that reported agreement for PDs across raters. Some of these studies featured a method in which two clinicians interviewed a patient jointly although most administered unstructured interviews separately. Samuel concluded that levels of agreement appeared slightly higher when the clinicians provided ratings with a standardized instrument (e.g., the SWAP), but were generally modest and only marginally within the range of agreement deemed acceptable by Kraemer and colleagues (2015). It is important to note, though, that in nearly all of these studies at least one of the interviews was administered by a separate clinician for the sake of research, not by someone involved in the clinical care of the patient. Very few studies have obtained PD ratings from two separate individuals who are both actively involved as a treating clinician (e.g., Hesse & Thylstrup, 2008). Clearly more research is needed on the naturalistic agreement between PD ratings by two clinicians who are providing clinical services to the same patient (e.g., group co-leaders; individual and group therapists; prescribing psychiatrists and psychotherapist).

Construct Validity. Clinicians are necessary for determining the interrater reliability of clinical diagnoses, and reliability is necessary for validity. Therefore, there is also a great deal of research that has employed clinician ratings as another source for testing hypotheses and investigating the validity of specific constructs. For example, Blais (1997) examined the hypothesis that PDs are linked with the domains of the FFM by obtaining ratings of the DSM-IV

PDs and the markers of the FFM from 100 practicing psychiatrists, psychologists, and social workers. Specifically, Blais concluded that this research extended prior findings by demonstrating that "these findings are not method or sample specific and thereby strengthens the conviction that these two personality systems are meaningfully related" (p. 391). A variety of other studies have utilized clinicians as a unique source of ratings to determine if prior findings (typically those from self-report questionnaires) generalize to these expert raters. Although there are exceptions (e. g., Blais & Malone, 2013), most of this research has suggested that clinician ratings do not fundamentally differ from self-report ratings in terms of structure (Morey, Krueger, & Skodol, 2013) or test-retest consistency (Samuel & Widiger, 2011).

Agreement with other Sources. A routine aspect of construct validation is the multitrait-multimethod matrix (Campbell & Fiske, 1959), which suggests that the same constructs should evince similar properties and correlate across methods. Building on this same idea are a number of studies that have examined the agreement of PD ratings by clinicians with those from other sources, such as self-report or semistructured interviews (Widiger & Boyd, 2009). Samuel (2015) provided a recent review of this literature and concluded that there was only modest agreement between PD ratings provided by treating clinicians and those from the other sources. The median dimensional agreement across the 27 studies that had reported such a correlation ranged from a .05 to .36, with an overall median of .23. This omnibus correlation did mask subtle, but important, differences across studies. For example, the review noted that clinician ratings agreed slightly more highly with PD diagnoses generated by semistructured interviews administered by research personnel (*mdn K = .30; mdn r* = .28) than with self-report questionnaires (*mdn K = .08; mdn r* = .22). Furthermore, when clinician ratings were aggregated using a systematic method, such as the SWAP, the agreement with other methods was enhanced

(*mdn r* = .33) relative to unsystematic methods (*mdn* = .19). Overall, though, this research has suggested that the PD diagnoses assigned routinely by clinicians in their clinical practice have relatively little overlap with the diagnostic ratings provided via semistructured interviews and self-report questionnaires in research settings. This has important—and potentially problematic—implications for the translation of empirical research into evidence-based practice. Indeed, if the individuals diagnosed with borderline PD (BPD) within a treatment study are decidedly different than those diagnosed with BPD in clinical practice, then the clinician cannot have confidence that the client will benefit from that empirically-supported treatment.

Comparative and Predictive Validity. Given the relatively limited agreement between clinicians' diagnoses and other methods, an obvious question that remains is "who is right?" Of course, the answer to any such question is rarely straightforward or clear-cut and the most likely scenario is reciprocal validity of alternate sources. This has certainly been the case for other types of informants, as they routinely increment self-report, but self-report also increments the informants (Carlson, Vazire, & Furr, 2011; Klein, 2003; Miller et al., 2005; Oltmanns & Turkheimer, 2009). Similarly, Hopwood et al. (2008) reported that semistructured interviews and self-report questionnaires have unique strengths for detecting specific diagnostic criteria for Borderline PD. It stands to reason, then, that there would be aspects of PD where self-report might be preferred (e.g., less observable mental processes, such as chronic feelings of emptiness), but other—perhaps more evaluative traits—for which clinician informants might be particularly valuable (Connelly & Ones, 2010).

To date, only a single study that has directly compared the reciprocal validity of clinician ratings to other sources. Samuel et al. (2013) examined the ability of PD ratings completed by clinicians, as well as self-reports and semistructured interviews to predict psychosocial

functioning over five years in a large clinical sample. The results of a series of hierarchical regressions indicated that clinicians' naturalistic diagnoses never provided incremental validity beyond the semistructured interviews and only incremented self-report questionnaires in one of the four comparisons. In contrast, the self-report and interview-rated PD diagnoses predicted significant variance in functioning beyond the clinician ratings in eight of the ten comparisons examined. Although, these were limited by the lack of a clinician-rated functioning variable, they represent a strong test of comparative validity of clinicians' naturalistic PD ratings. Future research that makes similar comparisons using a systematic method for clinician diagnosis would be highly informative in determining the relative validity of these sources.

Clinicians' Diagnostic Biases. In addition to the research examining the reliability and predictive validity of clinicians PD diagnoses, a sizeable literature has also examined their diagnoses for potential biases. It is well-known that clinical judgment is routinely inferior to statistical prediction due to the human mind's inability to accurately detect, encode, and weight all the information relevant to a given decision (Grove, Zald, Lebow, Snitz, & Nelson, 2000). Thus, it is not surprising that research has shown that practicing clinicians, as with any human facing a complex decision, often employ heuristics that can lead to a variety of biases when making PD diagnoses.

One such bias is the tendency for a lack of coherence between the overall diagnosis and the clinician's own ratings of individual criteria (Morey & Benson, 2016; Morey & Ochoa, 1989). In these studies, clinicians were asked to first provide a rating of the diagnostic criteria for all PDs and then to provide a separate rating of each PD as present or absent. Morey and Ochoa (1989) found routine inconsistencies between these methods as assigned diagnoses did not match the diagnoses that would be expected based on the criteria themselves in 72% of the 291 cases

rated. Major sources of the inconsistency appeared to be the overweighting of certain criteria that, when present, led to overdiagnosis, or were treated as exclusion criteria when absent. For example, borderline PD was often diagnosed among clients that exhibited the criteria of self-harm, self-damaging impulsivity, or affective lability even when a total of five criteria were not present (Morey & Ochoa, 1989). These findings were recently replicated using data from a DSM-5 field trial (Morey & Benson, 2016). Research by Kim and Ahn (2002) suggests that these departures from the criterion sets are due to clinicians' mental representations of PDs that view certain symptoms as more causally central to the diagnosis.

Research also suggests that clinicians alter their diagnostic impressions of case vignettes based on the demographic features of the client (Lopez, 1989). For example, gender bias for PDs has been examined in a number of ways (Widiger, 1998), but one common method has been to experimentally manipulate the gender of a case vignette and have clinicians provide diagnostic ratings (e. g., Warner, 1978). Although results have varied across studies, likely due to differences across the vignettes, the most common findings have been a tendency to diagnose histrionic PD more often in women and antisocial and/or narcissistic more often in men (Flanagan & Blashfield, 2005; Ford & Widiger, 1989; Samuel & Widiger, 2009). Potential biases on the basis of race and ethnicity are also a concern for the PD diagnoses (Garb, 1997). A few studies have also employed the case vignette methodology among clinicians to investigate possible racial biases. Interestingly, this research has suggested that clinicians within the UK are more likely to diagnose PDs for cases presented as white/Caucasian than as black/African-Caribbean (Mikton & Grounds, 2007).

In light of these findings it is relevant to consider the initial proposal of the DSM-5 Personality and Personality Disorders Work Group (PPDWG) was to abandon criterion sets in

favor a prototype matching approach (Widiger, 2011). An explicit impetus behind this proposal was to embrace the naturalistic diagnostic practices as "clinicians tend to think more readily in terms of prototype matching" (Skodol & Bender, 2009; p. 390). Although it may well be the case that clinicians naturally think in terms of prototypes, it is quite ironic to suggest that the official diagnostic procedures should bend in response (i.e., Shedler et al., 2011). Considering the demonstrated biases in these routine judgments due to these heuristics, it is much more logical to call for clinicians to changes their diagnostic practices than to accurately model that flawed approach. Fortunately, based on multiple critiques (e.g., Widiger, 2011; Zimmerman, 2011), this proposal was ultimately abandoned in favor of a much more empirically supported alternative. The path forward favors identifying the causes and consequences of diagnostic biases followed by training clinicians to utilize accurate and valid diagnostic approaches.

Limitations to the Validity of Clinicians' as an Assessment Source. Practicing clinicians provide PD diagnoses to individuals across the country every day and, as such, it is highly worthwhile to understand the cognitive processes by which those diagnoses are assigned, as well as determine their ultimate construct validity. In short, they represent a key focus of empirical research on PD diagnosis. Clinicians bring to bear a wealth of experience and a tremendous depth in the understanding of psychopathology. Nonetheless, the findings suggest that, no matter how well-trained or experienced, there are limitations to clinicians' PD diagnoses, just as there is for any single source. There are several reasons why clinicians' PD diagnoses may be limited.

First, their interactions with the client are confined to a single, highly-controlled context. Research suggests that clients' interactions within the consulting room are among the chief inputs for a clinician's diagnostic impressions (Westen, 1997). A therapy session is, nearly exclusively, a one-on-one situation in a specific setting with strongly-defined roles. As such, it

should not be surprising if a client's behavior is restricted compared other contexts. Second, clinicians typically interact with the client for only one hour per week, resulting in approximately ten hours of contact over a 12-week treatment. This obviously indicates that the therapist lacks direct access to the vast majority of a client's lived experience and so ultimately relies primarily on the client's report to fill in those gaps. Third, although some PD symptoms are manifest in terms of specific behaviors that can be directly observed, many more are inner feelings and states that must be inferred on the basis of self-report or by indirect observation. As is true for all human beings, this inferential process is inherently limited for complex judgments that rely on weighting multiple data points (Grove et al., 2000).

Importantly, these facts do not suggest that clinicians should be shunned as an assessment source, particularly when their ratings are aggregated systematically (Westen & Weinberger, 2004). Nonetheless, it does suggest that clinicians' PD ratings should not be considered an infallible gold-standard that some have suggested (Shedler et al., 2011). Clinicians, like all humans, have biases and these idiosyncrasies in the collection or aggregation of data should not be celebrated or reified. In sum, research on clinicians' mental processes and diagnoses are highly valuable for determining what actually happens during PD diagnosis, but they should not be interpreted as informing what *should* happen in PD diagnosis.

Clinical Utility

Another specific area where research with clinicians has been quite informative is in determining the clinical utility of the PD system. As was the case for past editions, clinical utility is explicitly emphasized in the DSM-5 as the preface indicates that "this edition of DSM was designed first and foremost to be a useful guide to clinical practice," (APA, 2013; p. xli). First et al. (2004) proposed that clinical utility should be defined as "the extent to which DSM assists

clinical decision makers in fulfilling the various clinical functions of a psychiatric classification system" (p. 947). These various functions include case conceptualization, communication with professional and lay audiences, differential diagnosis, choosing appropriate interventions, improving outcomes, and predicting future course (Mullins-Sweatt & Widiger, 2009). It is perhaps debatable to what extent the DSM should prioritize utility versus validity (Crego, Sleep, & Widiger, 2016). Certainly it is the case that even the most valid model would fail its purpose if it is not useful in clinical practice. Yet, it is also true that models that emphasize utility at the expense of validity will also fail, so some balance of these two is necessary. This goal of clinical utility, however, has proven largely elusive for the current personality disorder nomenclature (Verheul, 2005). Consequently, a relatively wide literature has emerged that has collected the perceived utility of various models of PD from practicing clinicians (Mullins-Sweatt & Lengel, 2012). This literature can mostly be divided into three primary methodologies: Ratings of prototypical cases or vignettes (Samuel & Widiger, 2006; Sprock, 2003), experimental manipulations (Glover, Crego, & Widiger, 2012; Rottman, Ahn, Sanislow, & Kim, 2009), and ratings of treated clients (Mullins-Sweatt & Widiger, 2011; Spitzer, First, Shedler, Westen, & Skodol, 2008).

Broadly, these studies have indicated much stronger support for the utility of the dimensional models relative to the DSM-IV categories. Further, when clinicians were provided with comparable methods and appropriate contextualization information, there did not appear to be appreciable differences in utility across a variety of dimensional models (Mullins-Sweatt & Lengel, 2012). A notable recent study replicated many of these prior findings using the specific components of the DSM-5 Section III PD model (Morey, Skodol, & Oldham, 2014). Morey and colleagues utilized a sample of practicing clinicians who rated their own clients in terms of this

model and then compared ratings of utility of those from the DSM-5 Section II model (i.e., the legacy categories). They found significant support in favor of the Section III system, with the strongest support for the dimensional traits, which were favored by psychiatrists and psychologists alike.

Finally, a study by Samuel and Widiger (2011) is notable as it remains the only one that has collected impressions of utility *after* clinicians had utilized the model longitudinally. Samuel and Widiger (2011) had clinicians provide diagnostic impressions for the FFM traits and the DSM-IV PD categories. Clinicians then provided ratings of clinical utility after the initial diagnostic ratings and then again after six months of treatment. Like prior studies, clinicians found the dimensional traits to be more useful than the categories. Importantly, though, the longitudinal nature of this study suggests that the traits were useful during the ensuing treatment. Nonetheless, this study and all others of clinical utility are ultimately limited in that they study *perceptions of* utility. An ultimate test will be to show that they are useful for improving outcomes and specifying the mechanisms by which various models provide information that enhances clinical decision making.

Issues to Consider when Sampling Clinicians

Considering all the available data we conclude that the sampling of practicing clinicians remains a valuable research practice, despite the acknowledged limitations. Specifically, care should be taken not to presume greater validity for clinicians' ratings. Rather, their diagnoses and ratings should be subjected to the same level of empirical scrutiny as any other method or source. With this in mind, there are several issues to consider when sampling clinicians.

Duration of Clinical Contact. A central consideration is how much clinical contact is required before a clinician can provide valid ratings. On the one hand, clinicians frequently

assign provisional diagnoses at the conclusion of an intake interview. On the other, given the complexity and potential stigma, clinicians are often understandably reluctant to provide a PD diagnosis based on a single meeting. Thus, the question is how long a clinician must know a client before they know the individual well enough to provide a reasonable diagnosis. Westen and Shedler indicate a minimum of six hours of clinical contact before completing the SWAP (Westen et al., 2012). Although a longer duration will certainly provide greater familiarity, it is complicated by the potential for clinical change due to therapy. Thus care should be taken to isolate initial diagnoses and current functioning from subsequent diagnoses.

Participation Rate. An inherent difficulty of research with clinicians is the rate of participation. Response rates to postal or electronic mail surveys typically hover around 10-15%, although some have been as low as 1% (Spitzer et al., 2008); with rates higher among psychologists than psychiatrists (Westen & Shedler, 1999). All too often, published articles do not even report a response rate. Given this, a primary limitation to collecting samples of clinicians involves the possibility of self-selection biases impacting findings. In an effort to increase participation, studies have offered honorariums to clinicians that approximate an hourly wage (e.g., up to $200/hour). Although there is no clear data on the impact this has on participation, it does appear that compensation is necessary for studies asking for than a few minutes from clinicians. An additional method of sampling, that has been used profitably in the ICD-11 revisions (Keeley et al., 2016), is to rely on practice panels of interested clinicians who are then surveyed for particular purposes. However, this type of approach likely reflects some inherent preferences and referral biases that again complicate findings in unknown ways.

Sample Size. In addition to relatively low response rates to mail surveys, it can be even more challenging to obtain large samples of practicing clinicians for more intensive studies. For

example, comparative studies that collect ratings from clinicians as well as report from the clients are particularly difficult. Most studies of that nature have samples that are quite modest. For example, Samuel (2015) reviewed 27 studies that compared clinicians PD diagnoses to other methods and the median sample size was 72. However, even that number is deceptively large as some studies simply recorded clinicians' chart diagnoses, requiring little effort from clinicians. For example, the five studies that employed the SWAP to collect clinician ratings had a median sample size of 47. This type of study is also quite expensive as collecting 50 clinician ratings might cost as much as $10,000 in direct participant payments. Given the current funding climate at NIMH focuses heavily on neurobiology, large-scale studies of this sort appear unlikely to find federal funding. Indeed, the most recent large-scale study of this sort was for data collected by Morey and colleagues on clinicians' use of the DSM-5 PD models and was "self-funded." It would be reasonable for the American Psychiatric Association themselves to fund this sort of research on the validity and utility of DSM diagnoses, but as yet, this has not occurred.

Regardless of the reasons, very few studies that have obtained large samples of practicing clinicians. What literature does exist typically relies upon passive participation (i.e., chart diagnoses) or makes do with smaller samples. Although statistical power should always govern sampling decisions from a scientific perspective, the practical limitations should also be recognized. Given the difficulty in obtaining the funding necessary for large scale diagnostic studies, researchers should continue to work toward creative solutions to systematically study clinicians' diagnostic impressions. One such option might be to integrate clinician ratings into treatment outcome studies. Not only would this provide a vehicle for more efficiently gathering therapist ratings, but it would also allow the direct comparison of therapist and self-report methods for predicting outcomes. A secondary option would be to organize practice networks

that integrate a standard assessment and outcome-tracking battery into clinical care across a broad sample of mental health practitioners. In the interim, if studies that employ practicing clinicians are held to the same sample-size expectations as studies that rely on self-reports or semistructured interviews, then the literature will remain sparse. Thus we see value of publishing findings from relatively small samples of clinicians -- perhaps as small as 40-50 – to build the literature on this important topic while recognizing the limitations of these sample sizes.

Method of Data Collection from Clinicians. A central question regarding research with clinicians is the method of data collection. Researchers seeking to employ clinicians face a difficult choice without appealing options. They must navigate between the Scylla of utilizing abbreviated measures comprise validity, but maximize sample size and the Charybdis of using well-validated, but longer measures that compromise participation from busy professionals. The vast majority of prior research has chosen the former and relied on brief, unsystematic ratings from clinicians. Yet an emerging literature suggests that these sort of clinician ratings lack validity, suggesting that more well-validated measures are needed to properly isolate the validity of this data source.

Conclusions

Given they are the front-line users of the diagnostic system, clinicians have much to recommend them as a potentially valuable source for research, such as through surveys of user acceptability or the study of the validity of their diagnostic impressions. Although there are practical obstacles, we hope this paper spurs greater interest in to integrating clinicians into PD research. Nonetheless, the existing research on the validity of their naturalistic diagnostic ratings also indicates that therapists have biases, just as with any assessment source, that cautions against the belief that they represent a gold-standard of diagnosis. Rather, our hope is that

continued research will determine how information from clinicians, as well as other sources, such as self-report, can be fruitfully integrated to improve diagnostic practice.

References

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders - Fifth Edition*. Washington, DC: Author.

Bernstein, D. P., Iscan, C., & Maser, J. (2007). Opinions of personality disorder experts regarding the DSM-IV personality disorders classification system. *Journal of Personality Disorders, 21*(5), 536-551. doi:10.1521/pedi.2007.21.5.536

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:10.1177/1073191112436667

Blais, M. A. (1997). Clinician ratings of the five-factor model of personality and the DSM-IV personality disorders. *Journal of Nervous and Mental Disease, 185*(6), 388-393. doi:10.1097/00005053-199706000-00005

Blais, M. A., & Malone, J. C. (2013). Structure of the DSM-IV Personality disorders as revealed in clinician ratings. *Comprehensive Psychiatry, 54*(4), 326-333. doi:10.1016/j.comppsych.2012.10.014

Block, J. (2008). *The Q-sort in character appraisal : encoding subjective impressions of persons quantitatively* (1st ed.). Washington, DC: American Psychological Association.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix. *Psychological Bulletin, 56*(2), 81-105. doi:10.1037/h0046016

Carlson, E. N., Vazire, S., & Furr, R. M. (2011). Meta-Insight: Do People Really Know How Others See Them? *Journal of Personality and Social Psychology, 101*(4), 831-846. doi:10.1037/A0024297

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:10.1037/A0021212

Crego, C., Sleep, C. E., & Widiger, T. A. (2016). Clinicians' Judgments of the Clinical Utility of Personality Disorder Trait Descriptions. *Journal of Nervous and Mental Disease, 204*(1), 49-56. doi:10.1097/Nmd.0000000000000424

Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed Self-Assessment: Implications for Health, Education, and the Workplace. *Psychological Science in the Public Interest, 5*(3), 69-106. doi:10.1111/j.1529-1006.2004.00018.x

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:10.1037/A0034878

Fiedler, E. R., Oltmanns, T. F., & Turkheimer, E. (2004). Traits associated with personality disorders and adjustment to military life: Predictive validity of self and peer reports. *Military Medicine, 169*(3), 207-211.

First, M. B., Bhat, V., Adler, D., Dixon, L., Goldman, B., Koh, S., . . . Siris, S. (2014). How Do Clinicians Actually Use the Diagnostic and Statistical Manual of Mental Disorders in Clinical Practice and Why We Need to Know More. *Journal of Nervous and Mental Disease, 202*(12), 841-844. doi:10.1097/Nmd.0000000000000210

First, M. B., Pincus, H. A., Levine, J. B., Williams, J. B. W., Ustun, B., & Peele, R. (2004). Clinical utility as a criterion for revising psychiatric diagnoses. *American Journal of Psychiatry, 161*(6), 946-954. doi:10.1176/appi.ajp.161.6.946

Flanagan, E. H., & Blashfield, R. K. (2005). Gender acts as a context for interpreting diagnostic criteria. *Journal of Clinical Psychology, 61*(12), 1485-1498. doi:10.1002/jclp.20202

Ford, M. R., & Widiger, T. A. (1989). Sex Bias in the Diagnosis of Histrionic and Antisocial Personality-Disorders. *Journal of Consulting and Clinical Psychology, 57*(2), 301-305. doi:10.1037//0022-006x.57.2.301

Ganellen, R. J. (2007). Assessing normal and abnormality personality functioning: Strengths and weaknesses of self-report, observer, and performance-based methods. *Journal of Personality Assessment, 89*(1), 30-40.

Garb, H. N. (1997). Race bias, social class bias, and gender bias in clinical judgment. *Clinical Psychology-Science and Practice, 4*(2), 99-120.

Glover, N. G., Crego, C., & Widiger, T. A. (2012). The clinical utility of the five factor model of personality disorder. *Personality Disorders: Theory, Research, and Treatment, 3*(2), 176-184. doi:10.1037/a0024030

Gritti, E. S., Samuel, D. B., & Lang, M. (in press). Diagnostic agreement between clinicians and clients: The convergent and discriminant validity of the SWAP-200 and MCMI-III personality disorder scales. *Journal of Personality Disorders*.

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

Hesse, M., & Thylstrup, B. (2008). Inter-rater agreement of comorbid DSM-IV personality disorders in substance abusers. *BMC Psychiatry, 8*. doi:10.1186/1471-244x-8-37

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.

Keeley, J. W., Reed, G. M., Roberts, M. C., Evans, S. C., Medina-Mora, M. E., Robles, R., . . . Saxena, S. (2016). Developing a Science of Clinical Utility in Diagnostic Classification Systems Field Study Strategies for ICD-11 Mental and Behavioral Disorders. *American Psychologist, 71*(1), 3-16. doi:10.1037/a0039972

Kim, N. S., & Ahn, W. K. (2002). Clinical psychologists' theory-based representations of mental disorders predict their diagnostic reasoning and memory. *Journal of Experimental Psychology-General, 131*(4), 451-476. doi:10.1037//0096-3445.131.4.451

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

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:10.1176/appi.ajp.2011.11010050

Lopez, S. R. (1989). Patient Variable Biases in Clinical Judgment - Conceptual Overview and Methodological Considerations. *Psychological Bulletin, 106*(2), 184-203. doi:10.1037//0033-2909.106.2.184

Mikton, C., & Grounds, A. (2007). Cross-cultural clinical judgment bias in personality disorder diagnosis by forensic psychiatrists in the UK: A case-vignette study. *Journal of Personality Disorders, 21*(4), 400-417. doi:10.1521/pedi.2007.21.4.400

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:10.1521/pedi.2005.19.4.400

Morey, L. C., & Benson, K. T. (2016). An Investigation of Adherence to Diagnostic Criteria, Revisited: Clinical Diagnosis of the Dsm-Iv/Dsm-5 Section Ii Personality Disorders. *Journal of Personality Disorders*, *30*(1), 130-144.

Morey, L. C., Krueger, R. F., & Skodol, A. E. (2013). The Hierarchical Structure of Clinician Ratings of Proposed DSM-5 Pathological Personality Traits. *Journal of Abnormal Psychology, 122*(3), 836-841. doi:10.1037/A0034003

Morey, L. C., & Ochoa, E. S. (1989). An investigation of adherence to diagnostic criteria: Clinical diagnosis of the DSM-III personality disorders. *Journal of Personality Disorders, 3*(3), 180-192. doi:10.1521/pedi.1989.3.3.180

Morey, L. C., Skodol, A. E., & Oldham, J. M. (2014). Clinician Judgments of Clinical Utility: A Comparison of DSM-IV-TR Personality Disorders and the Alternative Model for DSM-5 Personality Disorders. *Journal of Abnormal Psychology*, *123*(2), 398-405. doi:10.1037/a0036481

Mullins-Sweatt, S. N., Bernstein, D. P., & Widiger, T. A. (2012). Retention or Deletion of Personality Disorder Diagnoses for Dsm-5: An Expert Consensus Approach. *Journal of Personality Disorders, 26*(5), 689-703.

Mullins-Sweatt, S. N., & Lengel, G. J. (2012). Clinical Utility of the Five-Factor Model of Personality Disorder. *Journal of Personality, 80*(6), 1615-1639. doi:10.1111/j.1467-6494.2012.00774.x

Mullins-Sweatt, S. N., & Widiger, T. A. (2009). Clinical utility and DSM-V. *Psychological Assessment, 21*(3), 302-312. doi:10.1037/a0016607

Mullins-Sweatt, S. N., & Widiger, T. A. (2011). Clinician's judgments of the utility of the DSM-IV and five-factor models for personality disordered patients. *Journal of Personality Disorders, 25*(4), 463-477. doi:10.1521/pedi.2011.25.4.463

Oltmanns, T. F., & Turkheimer, E. (2009). Person Perception and Personality Pathology. *Current Directions in Psychological Science, 18*(1), 32-36. doi:DOI 10.1111/j.1467-8721.2009.01601.x

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

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

Regier, D. A., Narrow, W. E., Clarke, D. E., Kraemer, H. C., Kuramoto, S. J., Kuhl, E. A., & Kupfer, D. J. (2013). DSM-5 Field Trials in the United States and Canada, Part II: Test-Retest Reliability of Selected Categorical Diagnoses. *American Journal of Psychiatry, 170*(1), 59-70. doi:10.1176/appi.ajp.2012.12070999

Rottman, B. M., Ahn, W. K., Sanislow, C. A., & Kim, N. S. (2009). Can clinicians recognize DSM-IV personality disorders from five-factor model descriptions of patient cases? *American Journal of Psychiatry, 166*(4), 427-433. doi:10.1176/appi.ajp.2008.08070972

Samuel, D. B. (2015). A Review of the Agreement Between Clinicians' Personality Disorder Diagnoses and Those From Other Methods and Sources. *Clinical Psychology-Science and Practice, 22*(1), 1-19. doi:Doi 10.1111/Cpsp.12088

Samuel, D. B., Lynam, D. R., Widiger, T. A., & Ball, S. A. (2012). An expert consensus approach to relating the proposed DSM-5 types and traits. *Journal of Personality Disorders, 3*(1), 1-16. doi:10.1037/a0023787

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. (2004). Clinicians' personality descriptions of prototypic personality disorders. *Journal of Personality Disorders, 18*(3), 286-308.

Samuel, D. B., & Widiger, T. A. (2006). Clinicians' judgments of clinical utility: A comparison of the DSM-IV and five-factor models. *Journal of Abnormal Psychology, 115*(2), 298-308. doi:10.1037/0021-843X.115.2.298

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. (2011). Clinicians' use of personality disorder models within a particular treatment setting: A longitudinal comparison of temporal consistency and clinical utility. *Personality and Mental Health, 5*(1), 12-28. doi:10.1002/pmh.152

Shedler, J. (2002). A new language for psychoanalytic diagnosis. *Journal of the American Psychoanalytic Association, 50*(2), 429-456. doi:Doi 10.1177/00030651020500022201

Shedler, J. (2015). Integrating clinical and empirical perspectives on personality: The Shedler-Westen Assessment Procedure (SWAP). In S. K. Huprich (Ed.), *Personality Disorders: Toward Theoretical and Empirical Integration in Diagnosis and Assessment* (pp. 225-252). Washington, DC: American Psychologial Association.

Shedler, J., Beck, A. T., Fonagy, P., Gabbard, G. O., Kernberg, O., Michels, R., & Westen, D. (2011). Revision of the Personality Disorder Model for DSM-5 Response. *American Journal of Psychiatry, 168*(1), 97-98. doi:10.1176/appi.ajp.2010.10101466r

Skodol, A. E., & Bender, D. S. (2009). The future of personality disorders in DSM-V? *American Journal of Psychiatry, 166*(4), 388-391. doi:10.1176/appi.ajp.2009.09010090

South, S. C., Oltmanns, T. F., Johnson, J., & Turkheimer, E. (2011). Level of Agreement Between Self and Spouse in the Assessment of Personality Pathology. *Assessment, 18*(2), 217-226. doi:Doi 10.1177/1073191110394772

Spitzer, R. L., First, M. B., Shedler, J., Westen, D., & Skodol, A. E. (2008). Clinical utility of five dimensional systems for personality diagnosis: A 'consumer preference' study. *Journal of Nervous and Mental Disease, 196*(5), 356-374.

Sprock, J. (2003). Dimensional versus categorical classification of prototypic and nonprototypic cases of personality disorder. *Journal of Clinical Psychology, 59*(9), 992-1014. doi:10.1002/jclp.10184

Thomas, K. M., Wright, A. G. C., Lukowitsky, M. R., Donnellan, M. B., & Hopwood, C. J. (2012). Evidence for the Criterion Validity and Clinical Utility of the Pathological Narcissism Inventory. *Assessment, 19*(2), 135-145. doi:10.1177/1073191112436664

Verheul, R. (2005). Clinical utility of dimensional models for personality pathology. *Journal of Personality Disorders, 19*(3), 283-302. doi:DOI 10.1521/pedi.2005.19.3.283

Warner, R. (1978). Diagnosis of Antisocial and Hysterical Personality-Disorders - Example of Sex Bias. *Journal of Nervous and Mental Disease, 166*(12), 839-845.

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 I: Developing a clinically and empirically valid assessment method. *American Journal of Psychiatry, 156*(2), 258-272. doi:10.1521/pedi.2000.14.4.291

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.

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. (1998). Invited essay: Sex biases in the diagnosis of personality disorders. *Journal of Personality Disorders, 12*(2), 95-118.

Widiger, T. A. (2011). A shaky future for personality disorders. *Personality Disorders: Theory, Research, and Treatment, 2*(1), 54-67. doi:10.1037/a0021855

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. (2005). Evidence-based assessment of personality disorders. *Psychological Assessment, 17*(3), 278-287. doi:10.1037/1040-3590.17.3.278

Wood, J. M., Garb, H. N., Nezworski, M. T., & Koren, D. (2007). The Shedler-Westen Assessment Procedure-200 as a basis for modifying DSM personality disorder categories. *Journal of Abnormal Psychology, 116*(4), 823-836. doi:10.1037/0021-843x.116.4.823

Zimmerman, M. (2011). A critique of the proposed prototype rating system for personality disorders in DSM-5. *Journal of Personality Disorders, 25*(2), 206-221. doi:10.1037/a0022108.

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.