Stability and Change in Personality Disorders

Leslie C. Morey[1] and Christopher J. Hopwood[2]

[1]Department of Psychology, Texas A&M University, College Station, Texas 77843; email: leslie-c-morey@tamu.edu

[2]Department of Psychology, Michigan State University, East Lansing, Michigan 48824

Annu. Rev. Clin. Psychol. 2013. 9:499–528

First published online as a Review in Advance on December 10, 2012

The *Annual Review of Clinical Psychology* is online at http://clinpsy.annualreviews.org

This article's doi: 10.1146/annurev-clinpsy-050212-185637

Keywords

personality disorders, course, outcome, stability

Abstract

Stability is thought to be one of the major distinguishing features between personality disorders (PDs) and other forms of psychopathology. The development of more reliable PD assessments and the implementation of four major longitudinal studies on PD stability have provided critical data with which to evaluate the stability of PD features. Results from these and other studies reveal significant complexity in the interpretation of PD stability because of several issues that can impact stability estimates. Such estimates will vary as a function of the type of constructs being assessed, the type of stability being considered, the modality and reliability of the assessments being used, and the impacts of sampling. In this article, longitudinal research on PD stability is reviewed in the context of these issues. It is concluded that no single answer can be given to the question, "How stable are PDs?" and that future research and classification need to consider carefully and account for the complexity of this question.

Contents

INTRODUCTION

The fourth edition of the American Psychiatric Association's *Diagnostic and Statistical Manual of Mental Disorders* (DSM-IV-TR; Am. Psychiatr. Assoc. 2000) defines a personality disorder (PD) as "an enduring pattern of inner experience and behavior that deviates markedly from the expectations of the individual's culture, is pervasive and inflexible, has an onset in adolescence or early adulthood, is stable over time, and leads to distress or impairment" (p. 685). Since the introduction of the DSM-III in 1980 (Am. Psychiatr. Assoc. 1980), which placed PDs on a separate axis of diagnosis (Axis II), interest and research on PD have increased substantially (Blashfield & Intoccia 2000). Although there have been several changes in PD diagnosis, including in the number and types of disorders, stability has consistently remained a central tenet of the conceptualization of PD in every edition of the DSM, dating back to 1952 (Am. Psychiatr. Assoc. 1952).

The placement of PDs onto a separate axis in the DSM-III was designed to contrast these disorders with the episodic Axis I disorders, under the presumption that PDs were relatively more stable over time (Grilo et al. 1998). Studies of PDs before the introduction of the DSM-III, largely limited to antisocial or borderline phenotypes, tended to describe these disorders as stable in their presentation (Carpenter & Gunderson 1977, Grinker et al. 1968, Maddocks 1970, Robins et al. 1977). However, with the increase in research devoted to this topic since the DSM-III, several findings appear to raise questions about the presumption of stability, frequently indicating that some aspects of these disorders can in fact show appreciable improvement over time (Bateman & Fonagy 2000, Morey & Meyer 2012, Zanarini et al. 2010). Such findings suggest the need for some reconsideration of the centrality of stability for conceptualizing PDs.

In particular, the literature on PD course must be examined with respect to a number of issues that can affect interpretation of estimates of PD stability (**Table 1**). One basic issue involves defining the nature of the elements that comprise PD, which may include PD categories, PD symptoms, pathological traits, and indicators of psychosocial functioning. A second issue involves the type of "stability" being computed and conceptualized. Various aspects of stability are associated with different research questions and statistical approaches and can be influenced by different factors. They can also lead to rather different stability estimates. A third issue is that the approach used to assess PD features can affect stability. For instance, methods that are more sensitive to contextual dynamics, such as situation-based behavioral assessments, may show lower stability

Table 1 Conceptual and methodological issues in personality disorder (PD) research

Issue	Examples	Implications
Construct definition	Dimensions, categories, traits, disorders, and functioning	Dimensions almost always lead to higher estimates for PD reliability, stability, and validity. Dimensions should generally be preferred in stability research. It remains unclear how traits, symptoms, and functioning relate to one another, but some evidence suggests that traits and functioning are more stable than symptoms and disorders.
Type of stability	Differential, absolute, interindividual, structural, ipsative	Specificity is needed in defining the type of stability being considered. Deciding which type of stability to focus on determines the statistical approach and should be driven by the research question.
Instrumentation	Self-report, interview, informant report, behavioral performance, narrative	Assessment method may impact stability estimates. In general, multimethod assessment is optimal for providing a complete picture of PD stability and change.
Assessment reliability	Reliability, stability	The reliability of assessment measures impacts stability, but little is known currently about how to disentangle unreliability from substantive change. More frequent assessments are needed to understand these processes better.
Sampling	Clinical, student, community, children, aging	Younger samples tend to have greater rates of PD. Sampling individuals with higher scores on PD features will tend to lead to greater absolute change in PD features and potentially wider confidence intervals around estimates of other kinds of stability.

than those that tap more dispositional characteristics, such as questionnaires that inquire about long-term attributes. A fourth issue is that estimates of the reliability of such instruments vary appreciably; although limited reliability can theoretically lead to underestimates of stability using certain methods, distinguishing "unreliability" from "substantive change" can be quite complex. Finally, sampling issues can affect stability estimates. For instance, because personality stability is typically found to differ as a function of age, stability estimates for PD features may also depend, in part, on the age of the sample being considered. Differences are also commonly observed across clinical and nonclinical samples, most notably when the instruments used to select individuals into the clinical samples are also used in stability analyses.

In this article, we review recent research on PD stability in the context of these issues. Our goals in reviewing this research are to provide a contemporary statement about the stability of various features of PD, contextualize this statement with respect to important issues that can affect interpretation of PD stability estimates, evaluate stability as a necessary criterion for the diagnosis of personality pathology, promote a more nuanced and clinically useful conception of PD stability, and provide directions for future research. We begin by describing recent studies on the course of PDs, followed by a review of the impacts of issues in longitudinal research on stability estimates from the existing literature. We conclude by depicting general estimates for various forms of stability of PD conceptualized and measured different ways, with particular attention to how the current understanding of PD in the field relates to PD diagnosis and the upcoming fifth edition of the DSM.

LONGITUDINAL STUDIES ON THE COURSE OF PERSONALITY DISORDERS

Conclusions about the course of PDs are informed by research of varied types, including studies on the course of individuals with non-PD psychopathology (e.g., Durbin & Klein 2006), PD treatment studies (e.g., Leichsenring & Leibing 2003), and studies on the stability of normative personality traits (e.g., Donnellan & Robins 2009). However, we draw heavily from the results

of the following four recent longitudinal studies that were designed specifically to examine the course of PDs.

The Children in the Community (CIC; Cohen et al. 2005) study used a large random community sample of about 800 children from 100 residential areas sampled in two counties in New York. The study began in 1975 and is ongoing, making it one of the most ambitious prospective longitudinal studies conducted. In the CIC, the youngest participants were nine years old when PD assessment began. As is typical with studies of children in this age range, initial assessments made use of parent (maternal) reports as well as self-report of PD symptoms, although this eventually shifted to a strictly participant self-report in adulthood. Data collection for PD began in childhood, with a mean age of 14 (early adolescence), and participants were then reassessed in three subsequent waves of data collection with mean ages of 16 (mid adolescence), 22 (early adulthood), and 33 (adulthood).

The Longitudinal Study of Personality Disorders (LSPD; Lenzenweger 2006) was one of the first prospective longitudinal studies conducted to specifically target PDs (begun in 1990), and it had a specific goal of addressing the lack of empirical evidence regarding stability in PDs. The LSPD used a prospective multiwave panel design, with three assessment periods over the course of four years from 1990 to 1997 among roughly 258 college students drawn from undergraduate participant pools. An initial sample of 1,684 students was narrowed via screening criteria into two groups to ensure the representation of personality pathology: the Possible Personality Disorder group and the No Personality Disorder group. Membership in the Possible Personality Disorder group ($N = 129$) required an individual to meet diagnostic threshold for at least one PD based on either the structured interview or questionnaire, and membership in the No Personality Disorder group ($N = 121$) required meeting fewer than 10 PD symptoms.

The McLean Study of Adult Development (MSAD) (e.g., Zanarini et al. 2003, 2005, 2010) is unlike other studies reviewed here because of its specific focus on the course of borderline PD (BPD). The MSAD sample is composed of 290 hospitalized participants diagnosed with BPD using structured interviews and 72 comparison participants who met diagnostic criteria for at least one PD other than BPD. Since 1992, interview assessments of PD features and functioning as well as self-report assessments of traits have been conducted every two years. As of this writing, the MSAD has reported on waves of follow-up assessment carried out to 16 years (Zanarini et al. 2012).

The Collaborative Longitudinal Personality Disorders Study (CLPS; Gunderson et al. 2000, Skodol et al. 2005a) is a recently completed 10-year prospective, repeated measures study that targeted four specific DSM-IV-TR PDs as well as major depressive disorder (MDD) in the absence of PD as a comparison group. The original CLPS baseline sample of 668 patients consisted of patients assigned to a primary PD diagnostic group on the basis of a semi-structured diagnostic interview, with support from either a questionnaire or a clinician rating form. The diagnostic distribution was as follows: schizotypal PD, $N = 86$ (or 13% of the total sample); BPD, $N = 175$ (26%); avoidant PD, $N = 158$ (24%); obsessive-compulsive PD, $N = 154$ (23%). The MDD comparison group consisted of 95 patients (14%) who met interview criteria for current MDD, had no more than two criteria of any PD, and had fewer than 15 PD criteria in total. Participants in the CLPS study were assessed with interview and questionnaire measures of PD symptoms, traits, and functioning regularly throughout the course of the study.

ISSUES IN RESEARCH ON PERSONALITY DISORDER STABILITY

Although each of the aforementioned studies sought to clarify the course of PDs, they differed with respect to many of the methodological and conceptual issues described in **Table 1**. An overall theme of this review is that PD stability estimates tend to vary as a function of these issues. This

Table 2 A sampling of stability estimates from four longitudinal studies on the course of personality disorders

		Two years		Four years		Ten years	
		Differential	Absolute	Differential	Absolute	Differential	Absolute
PD category	INT	L kappa = 0.37[g]	M 35%[d] 30%[f]		L 55%[d]		L 93%[d] 85%[i]
	SR	L kappa = 0.38[g]	M				
PD dimension	INT	S r = 0.39[a] r = 0.54[b] r = 0.59[g]	S d = −0.02[b] d = −0.30[g]	S r = 0.39[a]** r = 0.49[b] r = 0.47[h]	M d = −0.17[a] d = −0.62[b]	M r = 0.39[a]*** r = 0.35[j]	L d = 0.60[a]*
	SR	S r = 0.67[b] r = 0.69[g]	S d = −0.18[b] d = −0.21[g]	S r = 0.64[b]	M d = −0.62[b]		
Pathological trait facets	SR	S r = 0.68[h]		S r = 0.68[h]**		S r = 0.56[j]	
Normative trait domains	SR	S r = 0.83[c] r = 0.74[h]	S d = 0.03[c]	S r = 0.78[c] r = 0.72[h]	S d = 0.06[c]	S r = 0.56[e] r = 0.68[j]	

Benchmarks are as follows: d = 0.4, 0.7 (Cohen 1992); r = 0.2, 0.4 (Cohen 1992); kappa = 0.40, 0.75 (Fleiss 1981); remission = 20%, 40%.
Superscripts indicate studies as follows: a, CIC (Johnson et al. 2000); b, LSPD (Lenzenweger 1999), c (Wright et al. 2012); d, MSAD (Zanarini et al. 2010), e (Hopwood & Zanarini 2010); f, CLPS (Grilo et al. 2004); g (Samuel et al. 2011); h (Morey et al. 2007); i (Gunderson et al. 2011); j (Hopwood et al. 2012).
Note: Interviews in the CIC were of both the youth and a parent. Categorical absolute stability is indicated by the proportion remitted at the follow-up assessment. Remission was defined as not meeting criteria for two years in the MSAD study and as not meeting criteria for one year in the CLPS study. Absolute stability for dimensions is keyed in the positive direction (negative for PD dimensions and positive for adaptive traits). Absolute stability estimates are based on the PD selected sample in the LSPD data. Dimensional absolute stability estimates for PDs are based on total symptoms in the CIC and LSPD samples and an average across individual PDs in the CLPS sample, because of reporting differences across studies. Absolute stability estimates were averaged across traits in the LSPD study. All results are averaged across individual PDs with three exceptions: in the MSAD, in which case results focus on borderline PD symptoms; in the CIC, in which antisocial PD was not assessed; and in the Gunderson et al. (2011) study, in which only the remission rate for borderline PD was reported. All stability coefficients assume the study baseline except for the LSPD two-year coefficients, which are based on intervals 2 and 3 because there was no assessment two years following baseline in that study.
Abbreviations: CIC, Children in the Community Study; CLPS, Collaborative Longitudinal Personality Disorders Study; INT, interview; L, large change; LSPD, Longitudinal Study of Personality Disorders; M, medium change; MSAD, McLean Study of Adult Development; PD, personality disorder; S, small change; SR, self-report.
*8 years.
**3 years.
***9 years.

can be seen in **Table 2**, which depicts stability estimates calculated different ways, across different intervals, in the CIC, LSPD, MSAD, and CLPS studies across two, four, and six years. Because **Table 2** summarizes much information, we briefly describe it here and refer to it throughout the review. Estimates in **Table 2** are given for both of the major forms of stability, absolute (mean-level group changes) and differential (retest correlations). Because different coefficients have been reported across these studies, several cells include multiple coefficients, whereas others include one or none. Different effect size estimates are given based on the type of stability and the type of scoring (e.g., categorical versus continuous). In order to more easily interpret the table, we also provide an indication of whether the effect would be regarded as small, medium, or large by previously published conventions (Cohen 1992, Fleiss 1981). One broad point that can be easily

observed in **Table 2** is that, whereas in some instances substantial change in PD is described, in others estimates suggest appreciable stability in such features. The following sections attempt to reconcile these apparently discrepant results by examining these issues in more detail.

Construct Definition

Dimensionality has been a fulcrum of recent debates on PD classification (Morey et al. 2000, 2007; Skodol et al. 2005b; Trull & Durrett 2005; Widiger 2007). Although PDs have been classified using a "categorical" model since their inception in the DSM-III, considerable research has been devoted to examining the utility of dimensional models in assessing for PD, and a shift in this direction in the DSM-5 appears imminent (Skodol et al. 2011).

It is critical to understand, however, that the dimensional-categorical distinction has multiple meanings. One meaning has to do with the representation of PDs as a collection of groupings into which a person may be assigned (albeit not mutually exclusive groupings). Each of these groups is basically conceptualized as a binary category (either it is present or absent), but they can also be represented in a more continuous manner in the form of symptom counts (Kass et al. 1985), prototype matching scores (Westen & Shedler 1999), or scores on dimensional traits (Clark 2007). This distinction can also refer to the possibility of an essential continuity of normal and abnormal personality functioning, where the latter is viewed as simply reflecting a severe extension of the former. Generally, the assumption of such continuity leads to a preference for dimensional trait models, whereas the assumption of discontinuity leads to a preference for categorical syndrome models. Notably, some have suggested that both continuous and potentially noncontinuous elements of personality pathology are important, leading to recommendations for hybrid models that contain both sets of elements (Hopwood 2011, Morey et al. 2007, Wright et al. 2012a).

A related issue involves basic assumptions about what features are considered to constitute a PD. Potential elements include symptoms (such as suicidal ideation), traits (such as impulsivity), and functioning indicators (such as vocational failures), all of which are represented to a greater or lesser degree in the DSM-IV criteria. Complications arise in the interpretation of PD stability because of the use of these varying emphases. For instance, longitudinal research sometimes reveals an empirical disconnection between the stability of PD features measured different ways or between the stability of various elements of PDs. The impacts of these issues on PD stability are reviewed in turn, beginning with a discussion of distributional assumptions, followed by a discussion of trait and disorder conceptions of PD, and concluding with a discussion of the role of patient functioning.

Distributional assumptions. A robust finding in research on PDs and other forms of psychopathology and personality is that reliability, stability, and validity coefficients are generally higher when variables are scored dimensionally rather than as discrete categories (Heumann & Morey 1990, Markon et al. 2011, Morey et al. 2007, Samuel et al. 2011). This pattern emerges clearly in **Table 2**, where stability estimates tend to be appreciably higher for dimensional PD and trait assessments than for categorical PD assessments. For instance, whereas two-year estimates for differential change in PD dimensions were small, estimates for changes in PD categories were large, even in a comparison of the same indicators in the CLPS sample (Samuel et al. 2011). Overall, these results suggest that one reason for recent findings of instability in PD features has to do with the use of categorical assessments of PD.

Such results indicate an appreciable loss of information when scaling PD features as categories rather than dimensions, which is consistent with hypotheses about the underlying dimensionality of PD features in nature (Clark 2007, Trull & Durrett 2005). Despite some initial evidence to

the contrary (Haslam 2003), contemporary taxonomic research generally supports the claim that most personality and PD variables are dimensional (Edens 2006, Edens et al. 2009, Haslam et al. 2012, Rothschild et al. 2003). One would expect that scoring natural categories as dichotomies would not seriously compromise assessment validity, assuming the assessment could reliably differentiate natural kinds. However, dividing a natural dimension into arbitrary categories would be expected to diminish assessment reliability and validity, thus leading to underestimates of stability. Inappropriate use of categorical designations and the associated implementation of essentially arbitrary diagnostic cut-scores can have other practical consequences as well, such as the calculation of empirically dubious prevalence statistics.

Traits and Disorders. As discussed above, the term "dimensional" can also be used to refer to assumptions about the degree to which PDs reflect behavior that is quantitatively continuous with normal behavior as opposed to behavior that is qualitatively different. For instance, a symptom count of PD features would be dimensional in the sense that individuals would not be placed into diagnostic categories, but it would still be categorical in the sense that the content of the assessment would focus upon symptoms that are thought to be discontinuous from normal personality functioning, with the resulting continuum representing the severity of these features. This, for example, is different from a dimensional perspective that holds that individuals with PDs demonstrate trait extremity that differs in degree, but not in kind, from the behavior of others.

A dimensional approach has the potential to more parsimoniously capture the features of PDs by depicting transdiagnostic dimensions that cut across different ostensible PD phenotypes, such as the impulsivity that is characteristic of both antisocial and borderline PDs (e.g., Clark 2007, Krueger et al. 2011). It also has the advantage of connecting PD taxonomy to a large literature on normative personality assessment, development, stability, and validity (Markon et al. 2005, Widiger & Simonsen 2005). Such a dimensional perspective is represented in the DSM-5 proposal, in which many of the 79 "symptoms" that serve as the diagnostic criteria of the DSM-IV PDs would be reconceptualized as potential indicators of 25 traits (Krueger et al. 2011) that themselves cohere into five higher-order factors resembling the normative Five-Factor Model of personality (Thomas et al. 2012, Wright et al. 2012d).

The more traditional medical model that has undergirded the DSM-III and DSM-IV conception of PD assumes that mental disorders are qualitatively distinct from normal behavior. In this approach, pathological symptoms comprise the content of PD categories, and there is no effort to systematically relate these symptoms to normative behavior. Potential advantages of this categorical model include ease of communication, familiarity to the mental health field, and links to a large body of research based on the syndromal approach—all of which have been cited to support retaining PD types in the DSM-5 (e.g., Paris et al. 2009, Rottman et al. 2009, Shedler et al. 2010, Zimmerman et al. 2012). In the DSM-5 proposal these elements would be retained through typological diagnoses that would be based on particular combinations of traits and functional deficits. However, in this proposal four of the DSM-IV PDs may no longer represent "official" types—schizoid, paranoid, histrionic, and dependent—rather, they could be considered to be "PD-trait specified," with particular traits (e.g., suspiciousness for paranoid PD or submissiveness for dependent PD) highlighted as the prominent personality features.

The DSM-5 proposal has the potential to take advantage of the strengths of both dimensional and syndromal approaches to PD. However, more research needs to be done on hybrid models that seek to integrate dimensional and categorical perspectives. Preliminary research evaluating the ability of trait dimensions to coalesce into reliable types has been mixed. Eaton et al. (2011) derived prototypes based on dimensional indicators of PD traits from several large and diverse

samples. Although they were able to derive an empirically defensible prototype model within each sample, the prototype model did not generalize across samples. This finding suggests that, even when types can be identified in a single sample, different samples are likely to suggest a different typology. This finding stands in contrast to the consistent finding of robust dimensions when personality data are subjected to factor analytic modeling techniques to derive dimensions (e.g., Markon et al. 2005, Thomas et al. 2012, Widiger & Simonsen 2005).

Although this result raises concerns about the DSM-5 proposal for trait-defined types, it is notable that the Eaton et al. (2011) results are based on trait assessments. It may not be surprising that an instrument developed to assess traits with links to normal personality would not necessarily lead to symptoms clusters that are discontinuous with normative features. This finding therefore does not preclude the possibility that certain symptoms that are not effectively measured by trait markers may indeed be nonnormative and even potentially discontinuous with normal functioning. A salient example is the difference between trait neuroticism, or the general tendency to experience negative emotions, and "cutting," a behavior that is commonly associated with borderline PD. Although neuroticism predicts self-harming behavior (Baetens et al. 2011, Mullins-Sweatt et al. 2012), many people with high scores on neuroticism scales do not cut themselves. Thus, a comprehensive assessment might include a measure of trait neuroticism as well as more specific indicators of pathological self-harming behavior.

Evidence for the differential stability of traits and symptoms. Testing the general hypothesis that trait and syndromal perspectives on PD provide important and incremental information would require focusing on what may be unique about traits and disorders rather than focusing on integration in a common model (Hopwood 2011, Morey et al. 2007, Wright 2011). Stability represents one potential point of distinction between traits and disorders. Longitudinal research suggests that syndromal disorders tend to be less stable than trait dimensions even when the disorders are scored dimensionally (Durbin & Klein 2006, Hopwood et al. 2012, Morey et al. 2007). For instance, in **Table 2** every stability coefficient for both pathological and normative traits suggests relatively small change. In contrast, several of the four- and 10-year stability coefficients for PDs reflect medium or even large change.

This pattern does not depict potential stability differences between traits and symptoms within disorders—differences that were an initial focus of longitudinal research on hybrid models. For example, Zanarini et al. (2003) reported that of the 81% of participants in the MSAD sample with self-harming behaviors or suicidal ideation at baseline, only 25% retained these symptoms by the six-year follow-up. However, such dramatic change was not evident with all symptoms. More specifically, borderline PD symptoms as measured in the MSAD study could be classified into two distinct areas, including "acute" symptoms that resolved relatively quickly over time and "temperamental" symptoms that were more stable and resilient to change. The former of these two types includes the symptoms of impulsive features, such as impulsivity and self-harm, whereas the latter group includes the affective and interpersonal features of borderline PD, such as anger, suspiciousness, and abandonment concerns. At 10-year follow-up, although only 12% of participants still met DSM-IV-TR criteria for BPD (Zanarini et al. 2007), roughly half the acute symptoms had decreased substantially (with less than 15% of participants retaining these symptoms). The temperamental symptoms decreased much less substantially (with 20% to 40% retaining these symptoms) (Zanarini et al. 2007).

McGlashan et al. (2005) extended these results to other PDs by examining stability differences in PD criteria using data from the first two years of follow-up in the CLPS. These results again suggested that some diagnostic criteria for PDs are more stable than others. For example, the most stable avoidant PD criteria were "feels socially inept" and "feels inadequate," and the least

stable was "avoids jobs with interpersonal contact." Such results are consistent with the possibility that PDs might be best conceptualized as hybrids of two elements: (*a*) stable personality traits that may have normal variants but that in PDs are pathologically skewed or exaggerated, and (*b*) dysfunctional behaviors that are attempts at adapting to, defending against, coping with, or compensating for these pathological traits (e.g., self-cutting to reduce affective tension, avoiding work situations involving many people because of shyness).

It should be noted that Morey et al. (2004) found in the CLPS sample that changes in the specific DSM-IV-TR criteria sets appeared to be quite internally consistent as a whole. Thus, although McGlashan et al. (2005) found that some BPD features demonstrated larger changes than others, the internal consistency findings indicate that changes in one BPD criterion tended to predict that other BPD features would change as well. Finally, Gunderson et al. (2011) found that data from a 10-year follow-up indicate that the nine BPD criteria had largely similar levels of decline with a similar rank ordering of prevalence as at baseline. Such results suggest that it is not clear that the DSM-IV-TR PD criteria can serve to cleanly differentiate stable traits from dysfunctional behaviors.

Fortunately, the CLPS project provided an assessment of normative and maladaptive traits in addition to assessments of DSM-IV-based PDs, which perhaps provides a more direct test of stability differences among different elements of personality pathology. Morey et al. (1999) described changes in Five-Factor Model traits observed in study PD groups across the first six months of the study, a period in which a number of seeming remissions were observed (e.g., Gunderson et al. 2003). These analyses demonstrated significant decreases in neuroticism in the borderline and obsessive-compulsive PD groups, increases in conscientiousness in the borderline PD group, and significant decreases in agreeableness in the avoidant PD group within the first six months of the study. However, it should be noted that the magnitude of these mean-level changes was not particularly large. For example, the largest mean-level change was that observed in neuroticism for borderline PD, which involved a 0.38 standard deviation decline calibrated against community norms (Costa & McCrae 1992). Thus, any significant mean-level changes observed would be considered to represent small effects by Cohen's (1992) convention.

However, as discussed below, differences in remission rates or group changes over time (absolute stability) such as those discussed above can be conceptually and empirically distinguished from differences in retest correlations (differential stability). Morey et al. (2007) reported the retest stability of both normative and abnormal traits at four-year follow-up, as well as for the DSM-IV-TR dimensional criterion counts as described previously. Although all of these indicators demonstrated statistically significant correlations over time, **Table 2** shows that the mean correlations for both the normative (0.72) and pathological (0.66) traits demonstrated greater stability than the DSM-IV-TR PD criterion counts (0.47). Likewise, the average 10-year stability correlations were 0.68 for the normative traits, 0.57 for the pathological traits, and 0.35 for the PDs (Hopwood et al. 2012; **Table 2**). The trait estimates did not significantly differ, but both were significantly larger than the estimates for PDs. This pattern persisted even after controlling for short-term reliability (an issue also discussed in detail below).

One might hypothesize that the lower stability associated with disorders is due to the relatively poor validity of PD assessment relative to the assessment of personality traits (Chmielewski & Watson 2009, Widiger 2005). However, validity research in the CLPS sample (Morey et al. 2007, 2012) shows that PDs and traits demonstrated similar levels of criterion-related validity using methodologically balanced (i.e., including both interview-based and self-reported criterion variables) outcomes. These results counter suggestions that stability differences between traits and disorders can be explained by unreliability in measurement, as psychometrically inferior scales should generally perform worse than superior scales in a predictive context.

The stability of functional impairments. Experiencing a remission in symptoms does not necessarily mean a full "recovery." In the MSAD (Zanarini et al. 2010), such a recovery was defined as possessing good social and vocational functioning while demonstrating a sustained reduction in diagnostic features (at least two years of no longer meeting study criteria for BPD). Although 93% of borderline participants had achieved symptom remission by the 10-year follow-up (**Table 2**), only 50% achieved this level of functional recovery. Furthermore, roughly one-third of these patients ultimately "lost" this recovery, through either a recurrence of symptoms or a decline in social or vocational functioning (Zanarini et al. 2010, 2012). The much higher rate of symptomatic improvement when compared to full recovery suggests that categorical diagnostic remissions in borderline PD are considerably more frequent than is the attainment of the level of psychosocial functioning needed to achieve a good global outcome. Given the suggestion of differences in the mean-level stability of some features of BPD relative to others, future investigations might be directed at studying the functional impact of the apparently more enduring affective and interpersonal features of the disorder.

Similarly, changes in PD features also did not necessarily lead to corresponding functional improvements in the CLPS study. Skodol et al. (2005c) examined the effect of a decrease in PD psychopathology on levels of functional impairment after one and two years. As with the MSAD, notable symptom reduction was observed in the CLPS sample over this time (Grilo et al. 2004, Samuel et al. 2011, Shea et al. 2002). However, for the sample as a whole, significant improvement in psychosocial functioning over time occurred in only three of seven domains: relationships with spouse or mate, recreation, and global social adjustment. In the case of impairment in recreation, romantic relationships, and global social adjustment, these improvements seemed largely the result of improvements in the MDD control group—meaning that groups of participants diagnosed with PDs at baseline tended to show little improvement in these areas. Improvements in symptoms and functioning can be compared directly from a dimensional perspective. Skodol et al. (2005c) observed an average improvement in functioning of $d = 0.12$ across indicators, whereas Samuel et al. (2011) reported an average symptom reduction of $d = 0.30$ (**Table 2**) during the first two years of the CLPS.

The Skodol et al. (2005c) study also examined the extent to which improvement in PD symptomatology could be linked to functional gains. It did not appear that improvement in PD led to increased social functioning, a finding that echoes the observation of improved psychopathology and persisting social dysfunction noted by McGlashan (1993) in his earlier study of BPD, and further suggested in the MSAD data. Gunderson et al. (2011) extended aspects of these investigations out to the full 10 years of the CLPS follow-up. With the repeated nature of the observations across years 2, 4, 6, 8, and 10, the functional improvement in different study groups did achieve statistical significance across functional categories. Nonetheless, the magnitudes of the improvements were limited to 0.30 standard deviations or less for individuals with baseline PDs.

Overall, certain principles can be derived from the pattern of evidence from longitudinal studies about the stability of different elements of PD. First, dimensional scores will routinely demonstrate stronger stability than categorical scores. Second, some elements of PDs may be more stable than others, and some PDs or traits may be more stable than others. Third, traits and functioning indicators are typically more stable than symptoms or disorder indicators.

Type of Stability

The existence of several conceptually and statistically distinct forms of stability also complicates discussions of PD stability and change (e.g., De Fruyt et al. 2006). This is particularly problematic because authors often use the general term "stability" without being explicit regarding the type to

which they are referring, as different types of stability can vary substantially. This can be seen in **Table 2**, in which the focus is on two kinds of stability, differential and absolute. When there are meaningful differences between these kinds of stability, absolute stability is generally lower than differential stability in PDs, meaning that symptoms tend to decrease over time on average, but the rank ordering of individuals within longitudinal samples remains roughly the same (Grilo et al. 2004). In the following section we describe and differentiate five types of stability[1]: differential, absolute, interindividual, structural, and ipsative.

Distinguishing differential, absolute, and interindividual stability. Of the five types of stability, differential and absolute stability are most commonly applied to PDs and thus are the focus of **Table 2**. As described in more detail below, differential stability refers to the rank ordering of individuals in a sample, and absolute stability refers to the average changes observed in a sample. Interindividual stability, a less commonly reported but nevertheless important coefficient, refers to variability around a group-level trajectory across individuals in a sample. **Figure 1** visually distinguishes absolute, differential, and interindividual stability. In each panel of **Figure 1**, four participants are assessed on some attribute (e.g., number of overall PD symptoms) at baseline and follow-up. In the first panel, the mean across these participants is constant, so absolute stability is perfect. However, rank ordering is completely different across assessments, leading to a retest correlation of zero. There is also meaningful interindividual instability since the trajectories for different individuals varied, with some individuals increasing and others decreasing. In the second panel, each participant declined by five points across assessments, exemplifying absolute instability. However, the rank ordering stayed exactly the same, and there was no variance in individual trajectories. Thus, the retest correlation is 1.00, and there is perfect differential and interindividual stability. In the third panel, all individuals start with similar values at baseline but fan out at the second assessment. There is no group mean change, so absolute stability is perfect. The retest correlation is 1, so differential stability is perfect. However, each person has a different trajectory, indicating interindividual-level instability.

Differential stability. Differential (also, rank-order or retest) stability refers to consistency in the rank ordering of individuals on a given trait over time. This type of stability is typically assessed with retest correlations over substantial intervals.[2] Higher retest coefficients indicate that individuals who are relatively high in that dimension at one point are also relatively high in that dimension at a second time point. Lower retest coefficients would suggest that baseline values are relatively ineffective predictors of the relative ordering of individuals at some follow-up assessment. Thus, differential stability coefficients could answer the question, How well will a PD assessment at baseline predict which people will have the most and least severe PD at a later time?

Overall, estimates from longitudinal studies of PD suggest moderate differential stability in dimensional assessments of PD features (e.g., 0.30–0.50) across extended intervals (e.g., 3–10 years) (**Table 2**). For instance, in the CIC study, children who were assessed on personality features were

[1]We note that a number of temporally sensitive methods have been effectively applied to study the dynamics of affects (Solhan et al. 2009), neurobiological processes (Herpertz et al. 2001), or interpersonal behavior (Russell et al. 2007) to understanding PDs. Because this kind of research is not about the stability of the diagnosis itself but rather the dynamics of individuals with the diagnosis, it is not reviewed here.

[2]Note that there are a number of ways to compute retest and absolute stability. We address this issue in two ways. First, to deal with potential differences in the computation of bivariate effects (e.g. r or ICC for differential stability), we apply effect size labels in **Table 2** that, although coarse, provide a general sense of differences across cells. Second, because many methods include other variables in model-based analyses such as cross-lagged or growth-curve models that can complicate interpretations across studies, we focus primarily on simple bivariate effects.

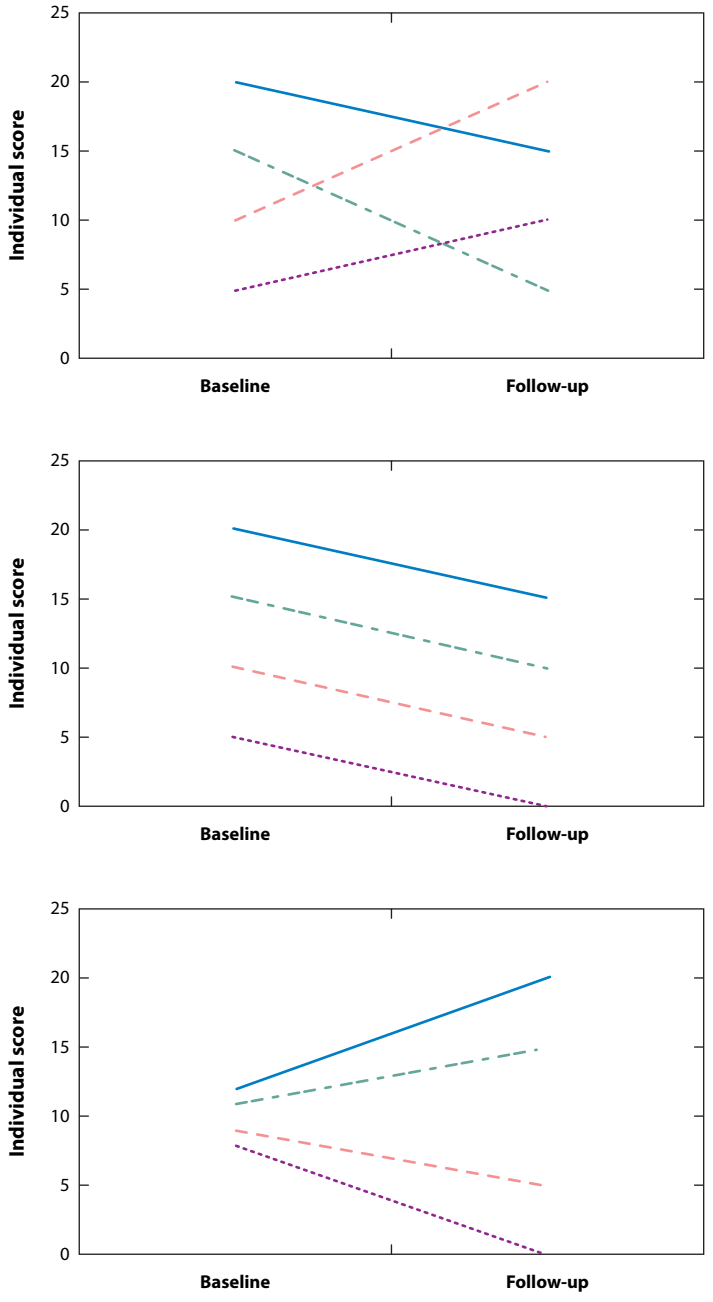

Figure 1

Differential, absolute, and interindividual stability. These figures reflect the scores of four hypothetical people on some attribute across two assessments. The score for each person is indicated by the y-axis, and each person is represented by a different line that indicates any changes in the person's score across assessments.

followed through adolescence and into adulthood. The retest correlation of a variable indicating the presence of PD in this sample averaged 0.40 over three years and 0.39 over eight years when using both youth and parent reports (Johnson et al. 2000; **Table 2**). The differential stability coefficient for total interview-assessed PD features in the LSPD data was 0.61, and the range of differential stability coefficients for individual PDs scored dimensionally ranged from 0.39 (paranoid) to 0.65 (antisocial), with an average of 0.49 (Lenzenweger 1999; **Table 2**). Estimates for individual PD dimensions in the CLPS averaged 0.86 at year 1 (Shea et al. 2002), 0.59 at year 2 (Morey et al. 2007, Samuel et al. 2011), 0.47 at year 4 (Morey et al. 2007), and 0.35 at year 10 (Hopwood et al. 2012) (**Table 2**).

Several other methodological issues discussed in this review need to be considered in interpreting these values, however. For instance, in the CLPS sample, 10-year estimates for PD stability increased when adjusting for assessment unreliability (mean = 0.47), and estimates for the stability of normative and pathological self-attributed traits in that study were considerably higher than those of interview-based PDs. As shown in **Table 2**, the 10-year retest correlation for PD dimensions was 0.39 in CLPS, whereas the 10-year retest correlations were 0.56 for pathological traits and 0.58 for normative traits (Hopwood et al. 2012). Also, the relatively low values in the CIC study were likely influenced by the use of nontraditional assessments as well as the age of the sample. For example, in **Table 2** the two-year retest correlation for PD dimensions in the CIC was 0.39 (Johnson et al. 2000), whereas the analogous correlation was 0.54 in the LSPD sample (Lenzenweger 1999) and 0.59 in the CLPS sample (Samuel et al. 2011). Thus, such methodological issues can potentially interact to further complicate efforts to estimate PD stability.

Absolute stability. Absolute (also, mean-level) stability refers to consistency in the average level of traits over time in a sample. Absolute stability for PD can be tested by comparing mean PD symptom counts or trait scores in the same sample over time, indicating the extent to which PD features are improving or worsening over time, on average. Absolute stability can also be approximated by tests of the proportion of individuals above or below a diagnostic cut-off at a follow-up assessment. Such values are a central consideration in treatment research, and evidence about absolute stability from the longitudinal studies reviewed below provides important information about the prognostic implications of a PD diagnosis for the typical patient.

Significant declines in PD features tend to be observed in treatment studies, longitudinal studies of individuals with PDs, and even community studies, although effects tend to vary across as well as within these kinds of samples. For instance, in **Table 2** it is clear that the absolute stability is higher in the CIC, a sample of individuals who were not selected for PD features, than in the LSPD, in which some participants were selected for PD features. These effects are likely to be even stronger in treatment samples. For example, Leichsenring & Leibing (2003) meta-analyzed a number of treatment studies to derive effect size estimates for change in PD patients treated with psychodynamic or cognitive-behavioral therapy. These studies showed that, although treatments were all conducted in one year or less, the Cohen's d effect size estimates for PD measures were large (mean of 1.46 for psychodynamic treatments and 1.00 for cognitive behavioral treatments), indicating that symptoms decline by a standard deviation or more, on average, in treatment studies. No effect sizes in **Table 2** reach that level, even when extended out to 10 years.

Instead, a significant but more modest decline was evident in the stability of the mean level of symptoms in nonclinical studies. Over four years in the LSPD (Lenzenweger 1999), the Cohen's d effect size for the subgroup selected for PD features was −0.62 (**Table 2**). Among nonselected participants, this effect size was −0.18. In the CIC, total PD traits declined 48% from an average of 16.6 symptoms when the sample averaged 9 to 12 years old to 8.6 when they were 25 to 28, which corresponds to a Cohen's d of −1.10. By the first follow-up period (two years after the

initial assessment) in the MSAD, 34.6% of BPD patients had remitted, with remission defined as no longer meeting interview criteria for BPD (Zanarini et al. 2003). Roughly half (49.5%) had remitted by four-year follow-up. By year six, 69% no longer met criteria for BPD (Zanarini et al. 2003), with 74% having remitted at some point over the entire course of the first six years. At 10-year follow-up, 93% of the initial BPD sample had attained a two-year period of remission from symptoms (Zanarini et al. 2007, 2010) (**Table 2**). By the sixteenth year of the study, nearly all patients in the study had experienced a remission, and symptom decline stayed relatively stable, with only a small proportion of the sample experiencing a return to diagnostic status (Zanarini et al. 2012). However, it is also notable that only about half of the borderline group and 75% of the other group had achieved significant functional improvements, and some had experienced relapse or worsened functioning (Zanarini et al. 2012). Over the 10 years of the CLPS, the large majority of patients no longer met full criteria for their baseline diagnosis (Gunderson et al. 2011). For example, using the stringent (12 consecutive months with two or fewer PD criteria met) definition of remission, roughly 85% of patients with borderline PD had remitted by year 10 (**Table 2**). These results indicate that long-term persistence of the features described in the diagnostic criteria are the exception rather than the rule, even in a sample with severe initial psychopathology, but that there are nevertheless some relatively enduring consequences to personality pathology in terms of functional difficulties.

One major conclusion from this research is that the remission rates for PDs found in longitudinal studies substantially exceed expectations derived from clinical assumptions as well as from prior long-term retrospective studies of outcome (McGlashan 1986, Plakun et al. 1985, Stone et al. 1987) and that these remission rates can be increased with treatment. However, it is important to note that reductions tend to occur primarily in the first parts of naturalistic studies (Lenzenweger 1999, Zanarini et al. 2007). These reductions may also be specific to certain kinds of symptoms, in contrast with other elements of PDs that tend to endure (Sanislow et al. 2009). Thus, research on absolute stability raises questions involving the effect of sampling individuals on the basis of high scores on instruments that are used to evaluate stability and the potential of such individuals to "regress to the mean" (discussed in detail below), in addition to the possibility that there exist in PD both enduring, trait-like elements and more dynamic symptoms as discussed above.

Overall, the disconnection between differential stability effects and absolute declines in longitudinal studies of PD is notable and informative. These findings suggest that although samples selected for PD features overall tend to experience substantial symptom reduction, there is considerable differential stability in the rank ordering of PD features across participants (Grilo et al. 2004). This pattern illustrates the importance of clarity with respect to the type of stability being conceptualized.

Interindividual stability. Interindividual (also, individual-level) stability refers to the homogeneity among individual PD scores over time. It provides an answer to the question, Do all individuals in the sample display a characteristic course that is well captured by the absolute stability coefficient, or is there significant variability around the average course? One method for detecting the presence of this kind of stability is to test for significant random effects around slopes in a growth-modeling context. In these models the slope term represents mean-level change for the entire sample. Variance in slopes confirms the presence of exceptions to the normative trend for the sample among particular individuals. For instance, whereas some individuals may have demonstrated dramatic increases or decreases, others could have remained fairly consistent over time. Although significant interindividual variability exists around the average reduction in PD symptoms observed in longitudinal studies (Hopwood & Zanarini 2010, Lenzenweger et al. 2004), little is known about what factors predict individual differences in growth trajectories (Lenzenweger

& Willett 2007, Wright et al. 2012b). One might hypothesize that factors such as treatment, the effectiveness of treatment, or significant life events might affect such trajectories. Such hypotheses represent an important topic for future research.

Structural stability. Structural stability refers to consistency in the patterns of covariation among traits. Structural stability is typically evaluated by testing the invariance of covariance matrices across measurement occasions. Structural stability is regarded as a prerequisite for examining absolute stability (Biesanz et al. 2003). However, relevant analyses are rarely reported, likely because the substantive clinical implications of structural variance over time are not obvious. Nevertheless, structural stability can have important clinical implications. For instance, Sanislow et al. (2009) found that the structure of the four PDs targeted in CLPS deteriorated over time. Whereas at baseline the disorders were quite distinct, at later intervals the diagnoses showed increasingly strong correlations with one another. This finding implies that the expression of PD is more specific when it is more severe (since overall rates of PD also decline over time), although this finding may also have reflected effects of the CLPS sampling strategy.

Ipsative stability. Ipsative stability refers to consistency in the patterning of personality traits within the individual. This type of stability captures configural changes over time in terms of profiles of indicators (e.g., traits or symptoms). In other words, investigations of ipsative stability address questions about the degree to which the constellation of attributes within the person is preserved over time. For instance, if an individual has predominantly paranoid as opposed to histrionic symptoms at baseline, will this pattern persist or will that person have more histrionic than paranoid symptoms a year later?

Ipsative stability tends to be higher in nonclinical samples than in clinical samples (De Fruyt et al. 2006), suggesting that instability in an individual's personality profile may be generally associated with pathology and personality immaturity (Hopwood et al. 2009). Hopwood and colleagues (Hopwood et al. 2009, Hopwood & Zanarini 2010) found that individuals with borderline PD tend to have greater levels of ipsative instability on personality trait profiles than do individuals with other PDs in both the CLPS and MSAD samples, consistent with theories positing identity instability as a core characteristic of this PD. These findings hint at the relevance of ipsative stability for understanding psychopathology and PD features and suggest a need for further work on this type of stability.

Instrumentation

The type of assessment instrument used to measure PD features can also impact stability estimates. A wide array of approaches is available for the assessment of PD features (Clark & Harrison 2001). These approaches differ with respect to assessment modality (for example, self-report versus observation versus interview versus informant ratings). The modality of assessing PD features is a critical issue in interpreting PD stability effects because the different kinds of instruments tend to yield different estimates of long-term stability.

Despite the wide availability of methods for assessing PD, the most common assessment methods in PD research are self-report questionnaires and diagnostic interviews (Huprich & Bornstein 2007). Thus, more is known about the impact of these methods on PD stability than is the case with other methods. In general, the differential stability coefficients of self-reported attributes tend to be higher than those of diagnostic interviews (Durbin & Klein 2006, Hopwood et al. 2012, Lenzenweger 1999, Samuel et al. 2011) (**Table 2**). For example, as shown in **Table 2**, Lenzenweger (1999) observed a differential stability correlation of 0.70 for total PD symptoms

on a self-report measure in the LSPD sample relative to 0.61 on an interview over four years. **Table 2** also shows that Hopwood et al. (2012) observed smaller 10-year differential stability of interview-based PD dimensions (mean = 0.35) relative to self-reported pathological traits (0.56) (see also Samuel et al. 2011). Consistent with these findings, Durbin & Klein (2006) reported 10-year stability coefficients for individuals selected for mood disorders. The average differential (intraclass) correlation for interview-assessed PDs was 0.49, whereas the average differential correlation for self-reported traits was 0.69.

There are several substantive reasons to expect that self-attributed characteristics might yield more differentially stable estimates than interview-assessed characteristics would. PD interviews require clinicians to ask individuals for specific behavioral examples of their symptoms and how they play out in day-to-day life. This frame of reference could magnify the importance of recent and more easily remembered experiences and thus emphasize more contextualized elements of behavior relative to the more context-free questionnaire items. This process may occur unevenly across patients or characteristics, leading to variability in the rank ordering of characteristics within people.

It is also possible that different assessment methods are more amenable to the measurement of more or less stable personality features, independent of their content. For example, interviewers might be generally inclined to attend closely to contextual factors that influence symptoms, despite instructions to consider enduring aspects of personality. Conversely, to the degree that it may be identity-reaffirming to see one's personality as basically stable, self-report methods could contribute to a level of consistency over time that overestimates objective trait consistency, given that self-reports are not direct measures of personality traits but rather measure the self-concept (i.e., how people see themselves) (McCrae & Costa 1994). Mischel (1968) famously offered such an interpretation, noting that "the trait categories people attribute to themselves and others may be relatively permanent, and may be more enduring than the behaviors to which they refer" (p. 36).

It is also possible that interviews and questionnaires will vary in their sensitivity to different aspects of personality and related pathology content. For example, Hopwood et al. (2008) found that a diagnostic interview and a self-report questionnaire for borderline PD that were matched in terms of item content were similarly valid for predicting five-year functional outcomes, but also that each had unique strengths. Specifically, the interview method for assessing more observable behaviors (e.g., impulsive or self-harming behavior) better predicted functional outcomes than did the use of self-report to assess the same features; conversely, the self-report measure of more inferential symptoms (e.g., emptiness, identity problems) had better predictive validity than did the interview approach. Similarly, McGlashan et al. (2005) reported that impulsive behaviors were among the most stable criteria of BPD as assessed by interview, and identity problems were among the least stable. Interestingly, in the CIC study, somewhat higher differential stability coefficients were observed when participant self-report was used than when estimates were based on combined parent and child reports (Johnson et al. 2000). Thus, it is possible that interviewers, parents, and others provide less reliable and valid ratings of some aspects of PD (e.g., those that require more inference) than can be provided by the person who directly experienced the feature at question.

One complication in reviewing this research is that assessment method has often been conflated with item content. Specifically, self-reported traits are often compared with interview-diagnosed PDs. It is difficult in such designs to determine whether stability differences across these assessments have to do with the nature of the variables being investigated (trait or disorder) or the method used to investigate them (questionnaire or interview). More research in which assessment method and content are compared directly is therefore needed to clarify the impact of interview

and self-report methods on the stability of PD features. As it stands, basic issues about the relative strengths and weaknesses of each approach are poorly understood. More broadly, longitudinal research using other methods, such as informant reports (Klonsky et al. 2002), laboratory tasks (Durbin et al. 2009), narrative data (McAdams et al. 2006), performance-based approaches (Mihura et al. 2012), and assessments taken in more ecologically valid settings (Solhan et al. 2009), will be critical for a nuanced and accurate understanding of the impact of assessment modality on PD stability.

Assessment Reliability

A basic issue related to assessment involves reliability and its association with stability. Classically defined as freedom from measurement error (e.g., Guilford 1954), reliability represents a conceptual upper bound for the validity or stability of personality pathology. The introduction of diagnostic criteria into the DSM-III in 1980 was explicitly intended to increase the interrater reliability of psychiatric diagnosis, and its introduction was followed by a considerable increase in the construction of standardized PD assessments designed to further improve reliability (Loranger et al. 1991, Zimmerman 1994).

As opposed to interrater reliability, retest reliability (Grilo & McGlashan 1999, Zanarini et al. 2000, Zimmerman 1994) is closely linked conceptually to differential stability. In fact, retest reliability and differential stability are typically computed using the same analytic methods and are only conceptually distinguished by the interval between assessments. Evidence that PD symptoms are subject to change, even over relatively brief intervals, makes it difficult to disentangle retest reliability from differential stability. But how does retest reliability affect absolute stability?

Consider a study by Loranger et al. (1991) that found substantial decreases in symptoms for all but two DSM-III-R (Am. Psychiatr. Assoc. 1987) PD diagnoses between 1 and 26 weeks following baseline measurements in a clinical sample. Was this due to significant symptom reductions or unreliable baseline assessments? Some authors note that reliability issues could affect not only differential instability but absolute instability of the kind observed by Loranger et al. (1991) as well. For example, Chmielewski & Watson (2009) noted that "due to the nature of clinical samples, transient error could influence mean-level change and remission rates as well as rank-order stability" (p. 200).

Generally speaking, retest reliability is thought to assess the influence of transient error reflecting fluctuations in an individual's psychological state on a given day (Green 2003, Schmidt et al. 2003). However, with a time interval of six months, as in Loranger and colleagues' (1991) study, it is difficult to determine whether the low stability rate is due to measurement error, true change, or a combination of both. Cattell et al. (1970) used the term "dependability" to conceptualize the consistency of variables during an interval in which substantive change is unlikely, which is essentially test-retest reliability over a brief period of time. Chmielewski & Watson (2009) suggested that, because PD characteristics are unlikely to change over the course of two weeks, two-week retest intervals could be used to estimate reliability, and longer-term differential stability coefficients could be corrected for these retest coefficients.

It is worth considering more carefully the potential link between retest correlations indicating differential stability over a short interval and absolute stability. Chmielewski & Watson (2009) suggested that transient error could invalidate baseline assessments in the direction of increased false positives, which would qualify at least some observed remissions as potentially false as well. However, research in naturalistic studies has not borne this prediction out. As discussed above, early results from the CLPS project revealed that fewer than half (i.e., 44%) of PD patients

remained at or above full criteria every month within the first year of follow-up (Shea et al. 2002). Because these results were surprising in light of the presumption that the identified features of PD should persist across much of adult life, Gunderson et al. (2003) closely examined the characteristics of 18 patients initially diagnosed with BPD who demonstrated a dramatic "remission" of these features—defined as presenting with fewer than two BPD criteria at reevaluation—over the first year of follow-up. In only two of these cases was the initial BPD diagnosis deemed possibly invalid (however, for an alternative view, see Widiger 2005). Instead, in most cases these "remissions" were associated with improvements in comorbid Axis I disorders or with resolution of situational crises. Thus, the conclusion from examining this subgroup of patients was that in some patients, the features of borderline PD can be more dynamic than presumed—in these cases being particularly sensitive to changes in the environment or in the clinical picture.

Furthermore, previous results from the CLPS data suggest that trait and disorder assessments are not differentially affected by baseline mood, which was one mechanism for diagnostic unreliability posited by Chmielewski & Watson (2009). Morey et al. (2010) critically tested this issue by comparing individuals diagnosed with PD in CLPS who either did or did not meet criteria for comorbid major depression at baseline. If mood issues invalidated the assessment of PD at baseline, those with depression should exhibit a more rapid decline in symptomatology and an increase in functioning with the remittance of the putatively less stable mood problems. However, those PD patients with comorbid depression at baseline functioned similarly six years later to those without comorbid depression, and both functioned much worse than MDD patients who did not present with PD symptoms at baseline. These results support the contention that "PD diagnoses established during depressive episodes are a valid expression of personality pathology rather than an artifact of depressive mood" (Morey et al. 2010, p. 528).

Research also calls into question the claim that addressing issues of reliability would bridge the gap in stability estimates of traits and disorders (discussed below). Hopwood et al. (2013) calculated 10-year dependability-corrected differential stability estimates for both traits and disorders in the CLPS sample. Mean stability estimates for the 10 PDs, corrected for reliability attenuation, were 0.47 using the test-retest reliability estimates, relative to 0.70 for pathological traits and 0.74 for normative traits. Furthermore, the PDs for which participants were selected into the CLPS study (borderline, avoidant, obsessive-compulsive, and schizotypal) had stabilities (average dependability-corrected stability = 0.49) similar to the PDs that were not selection criteria (0.45) in the study. If regression to the mean were largely responsible for the observed appreciable remission rates, it would be anticipated to affect selected PDs relatively more than nonselected PDs. Thus, although it is clear that adjusting stability coefficients for short-term assessment reliability will tend to increase those coefficients, it is not clear that short-term retest reliability differences are responsible for observed patterns of absolute change, remission, or the differences observed in stability estimates of traits and disorders in longitudinal studies of PD. Instead, retest reliability is most likely to affect differential stability, which tends to be relatively high in these studies, even over fairly long intervals.

However, there is much more to learn about the complex relationship between various forms of reliability and stability. More evidence-based approaches to modeling situation-specific, trait, and autoregressive components of differential stability in PD features (e.g., Anusic et al. 2012) are currently being developed and applied to personality data. Such approaches are likely to be informative about where to draw nonarbitrary lines between retest reliability and differential stability and to more precisely decompose potential influences on PD stability. Until sampling strategies assess PD features at frequent enough intervals to empirically address these issues, however, long-term differential stability estimates need to be interpreted cautiously in light of the potential influence of measurement unreliability.

Sampling

A final critical methodological context for interpreting stability estimates involves sampling. Obviously, longer sampling intervals will generally result in lower stability coefficients (Roberts & DelVecchio 2000). Indeed, within rows, coefficients systematically decline over time in **Table 2**. As a specific example, absolute changes in PD symptoms are classified as "small" according to Cohen's conventions at two years, "medium" at four years, and "large" at 10 years. However, the composition of the sample may also be impactful. For instance, issues such as age, clinical status, and demography may all moderate various types of PD stability.

Age. The assumption that a PD is stable and enduring implies that a PD should be apparent at a young age (DSM-IV) and be maintained throughout adulthood and into old age. However, the subject of PDs in youth is controversial (Krueger & Carlson 2001), in that the full-range of symptoms needed for a diagnosis of PD under the current DSM's classification system is difficult to observe in childhood and early adolescence. It is particularly difficult to infer that "personality-related" problems such as shyness or aggressiveness evident during childhood and adolescence will persist to adulthood (e.g., Loeber & Dishion 1983), although it should be noted that adult PDs will not necessarily persist, either (Grilo et al. 2004).

Childhood and adolescence are obviously salient periods for investigating the development of PD, given the apparent role of temperamental and developmental experiences in PD etiology. In addition to potential influences of temperament (Crawford et al. 2001, Rettew et al. 2003) and genetic dispositions (e.g., Distel et al. 2010), there is considerable evidence, particularly from the CIC study, suggesting that adult PDs are linked to early experiences (Cohen 1996, Cohen et al. 2005, Crawford et al. 2001, De Clercq & De Fruyt 2007). The correlations between traumatic experiences in childhood and adult PDs are well established (Cohen et al. 2001), although it is possible that these correlations could reflect gene-environment associations or other such factors (Bornovalova et al. 2012). Overall, to date relatively little is known about disordered personality in children and adolescents, underscoring the need for additional study (De Clercq & De Fruyt 2007, Krueger & Carlson 2001, Shiner 2009).

One general conclusion from existing research, however, is that PD levels tend to decline as individuals mature from childhood to adulthood. The results from the CIC study suggest that mean levels of PD symptoms in the community are generally highest in early adolescence and steadily decline into adulthood (Johnson et al. 2000). By the mid to late twenties, it appeared that mean levels of PD symptoms in this community sample stabilized considerably. This pattern is generally consistent with research on normal personality traits and other trait-linked psychopathology (Donnellan & Robins 2009). In nonclinical studies of normal personality, average trait levels tend to change toward increasing maturity and health (i.e., decreasing neuroticism; increasing extraversion, agreeableness, and conscientiousness) during adolescence and then become fairly stable during adulthood (Donnellan & Robins 2009, Roberts et al. 2006). Thus, the greatest normative risk for most forms of PD is during adolescence and early adulthood.

These findings raise the possibility that PD features track developmentally with changes in dispositional traits. This hypothesis was supported in a study by Wright and colleagues (2012c), who showed in an LSPD sample that reductions in avoidant PD symptoms track together over time with increases in dominance and warmth and decreases in neuroticism. These results are consistent with the hypothesis that normative traits reflect a core component of PD, as suggested by dimensional and hybrid conceptual models.

As with childhood PD, detecting personality pathology in aging samples can be challenging within the confines of the current DSM classification system (Oltmanns & Balsis 2011). For

example, it can sometimes be difficult to distinguish personality pathology from distress caused by the aging process itself (Abrams & Bromberg 2006). As an illustration, in the case of disabled or partially disabled older adults, there could be difficulty in differentiating dependency based on personality from dependency based on circumstance. In general it appears that the elderly tend not to experience fluctuations in personality traits over time (Roberts & DelVecchio 2000). However, it has also often been observed that certain disorders, such as antisocial PD or BPD and its associated impulsivity (Paris & Zweig-Frank 2001, Stevenson et al. 2003), decrease in severity as individuals grow older. Conversely, PDs involving eccentricity and anxiety can worsen with age (Seivewright et al. 2002). Thus, the stabilization of personality traits over time does not necessarily correspond to increased stability of PD diagnosis for the aging or to a more successful outcome. Future research, such as Oltmanns's St. Louis Personality and Aging Network study (e.g., Powers & Oltmanns 2012), that focuses on the development of PD features in aging populations will thus be highly informative.

Clinical status. A major issue in PD research involves the implications of findings obtained from samples that are unselected (student or community) as opposed to selected (based on diagnosis or from clinical settings) for an elevated rate of PD features. It is typically easier and more cost-efficient to sample and subsequently follow individuals from the community, particularly college students, than it is to sample patients. One advantage of sampling college students in addition to convenience is that, as discussed above, developmental trends in personality (Arnett 2000, Donnellan & Robins 2009, Roberts et al. 2006) suggest that PD features can be expected to be somewhat elevated and to decline somewhat more sharply relative to older community samples, similar to what would be expected in clinical samples but of a lesser magnitude.

However, the overall lower levels of personality pathology that can be expected in students and individuals randomly sampled from the community relative to patients can lead to range restriction and thus diminished validity and stability correlations, particularly for instruments with items that are not adapted to assess subthreshold features thoroughly, such as diagnostic interviews. On the other hand, when the instruments used to assess PD have a similar amount of reliable variance in selected and unselected samples and strong and specific convergence with diagnostic interviews in clinical samples, one can generally expect a similar pattern of validity and differential stability correlations in nonselected and selected samples. This is important because it suggests that researchers and clinicians can draw upon research using such populations to better understand PDs, although this is not to say that student samples should be regarded as a completely effective substitute for samples selected for clinical features.

In the end, the most direct answer to questions such as "How quickly will this patient's symptoms remit?" or "Which of these patients will be doing better a year from now?" logically comes from studies of patients as similar to those being considered as possible. Therefore it has traditionally been assumed that the best way to ensure generalizability and applicability of research results to patients is to conduct research on patient samples.

Although clinical samples are generally ideal for research on clinical problems, there are potential issues associated with studying PD stability in patient samples as well. This is particularly the case when the instrument that is used to select individuals into the study is also used to assess stability. First, this can restrict range on the variable of interest, which could actually lead to diminished validity and stability estimates in clinical samples. Furthermore, there is a risk of regression to the mean in studies using repeated measures with screened psychiatric patients, which can naturally cause an apparent decline over time. This is an issue in treatment studies as well as naturalistic longitudinal studies that sample individuals on the basis of measures and constructs whose stability

was an explicit focus of the study. Notably, three of the four major contemporary studies reviewed on PD stability in detail here (all but the CIC) sampled based on PD characteristics.

As discussed above, the LSPD (Lenzenweger 2006) sampled college students drawn from undergraduate participant pools, but oversampled individuals with PD features, thus having some of the advantages of both unselected and selected designs. The LSPD stability coefficients found by Lenzenweger (1999) are comparable to other longitudinal studies of PD using different samples and different measures. For example, the two-year differential stability effects reported in **Table 2** show that the LSPD stability coefficient is more similar to that of the CLPS sample (Samuel et al. 2011) than that of the CIC (Johnson et al. 2000). However, the effect of sampling on PD features on the stability of those features was apparent in the LSPD. Whereas the four-year Cohen's d for change in total PD symptoms by interview was −0.62 for the group selected for PD features in that study, that coefficient was −0.18 for the comparison group (Lenzenweger 1999). Such results point to the possibility that the PD-selected group regressed to the mean over time above and beyond developmental changes in the direction of increased maturity that may have impacted both groups.

Nevertheless, focused analyses in CLPS suggest that the sizable percentage of remissions observed in that study are not an artifact of unreliable assessment (i.e., the mechanism resulting in regression to the mean). Grilo et al. (2004) used the estimated test-retest reliability of criterion counts in CLPS diagnostic interviews (Zanarini et al. 2000) to estimate the number of criteria that would be observed in a subsequent interview given the number of criteria observed at baseline. These reliability-adjusted criterion counts were retransformed into raw criterion counts to allow prediction of the presence or absence of diagnosis. Those analyses revealed that the mean expected number of criteria accounting for reliability regression, while lower, was generally not sufficiently low to bring participants below the stringent definition of two or fewer criteria needed for remission. In fact, predicted remission rates due to unreliability ranged from 1% to 4%. Nevertheless, the ideal study would sample individuals using a different methodology than was the basis for selection, given the potential of regression to the mean to cause overestimates of remissions or absolute declines.

Finally, clinical samples not only select individuals based on PD features, they also tend to draw from treatment settings that implement interventions designed to impact absolute stability/remission rates. Thus, faster declines can generally be anticipated in treatment as opposed to other kinds of samples, as discussed above. This not only applies to patients enrolled, for example, in a clinical trial, but also to individuals in naturalistic clinical samples such as the CLPS and MSAD. Indeed, most CLPS patients were in treatment at the time of the baseline assessment, and there was considerable variability in subsequent treatments received by these patients (Bender et al. 2001). However, remissions continued steadily across the 10-year course of the study (Gunderson et al. 2011), even in the absence of sustained treatment (Bender et al. 2001, 2006). As shown in **Table 2**, 93% of CLPS patients diagnosed with BPD had remitted from that disorder by the 10-year follow-up. Furthermore, there appeared to be no significant effects of treatment intensity on the stability of PD criteria in CLPS patients (Grilo et al. 2004). This pattern may reflect the tendency for problem severity to drive the amount of treatment received in naturalistic studies (Cochran 1983). The apparent disconnect between treatment and naturalistic studies in terms of absolute stability, then, may be attributable to the varying treatment doses across individuals sampled or to the delivery of systematically followed and evidence-based treatments in treatment studies.

CONCLUSIONS

It is clear from this review that it is not possible to provide a single estimate for "the stability of personality disorder." First, research consistently suggests that some elements of PDs are more

stable than others. Second, a host of issues affect the interpretation of any obtained PD stability estimate. In reviewing a number of such issues it is apparent that each of them can, under certain circumstances, affect PD stability. This review provides an important framework within which to consider issues of stability as applied to PD diagnosis and classification and to provide a context for directions for future research. For instance, it can generally be anticipated that stability will be lower when variables are scored categorically rather than continuously. In general, traits and functioning can be expected to be more stable than symptoms, questionnaires more stable than interviews, and differential stability estimates to be higher than absolute stability estimates. When samples are younger, more severe, or in treatment, they are more likely to change than otherwise, and change is particularly likely to be observed when individuals are selected into studies using the same features that will be evaluated for stability.

Stability as a Criterion for Personality Disorder Diagnosis

The effects of numerous factors on stability estimates indicate that the longstanding characterization of PDs as enduring, stable, and beginning in adolescence lacks nuance. Doubts also arise when stability coefficients observed for PDs and other clinical disorders are compared. For example, Shea & Yen (2003) contrasted the diagnostic stability of PDs and functioning in the CLPS with the stability of anxiety, depression, and functioning in the Collaborative Depression Study (Katz & Klerman 1979) and the Harvard/Brown Anxiety Research Program (Keller et al. 1994). They observed that "PDs are less stable, and anxiety disorders are far more stable, than presumed by the conceptualizations of Axis I and II disorders" and that "stability does not provide a meaningful distinction" between these forms of psychopathology. They also found that "personality disorders do not have more persistent functional impairment than Axis I PDs" (Shea & Yen 2003, p. 379). Krueger & Markon (2006) similarly concluded that "stability is not an especially compelling differentiator of PDs and clinical disorders" (p. 23).

As of this writing, the DSM-5 proposal to define PDs indicates that the features of such disorders are "relatively stable across time and consistent across situations" (**http://www.dsm5.org**). This relative stability applies to both the observed impairments in personality functioning as well as the individual's personality trait expression, reflecting an integration of features of traits and disorders. In other words, DSM-5 PD types are defined by trait constellations (e.g., Blashfield 1993, Livesley 2007) in combination with core impairments in personality functioning specific to those types (Skodol 2012). This strategy implements suggestions from research regarding the importance of incorporating traits (Krueger et al. 2011, Samuel & Widiger 2008, Watson et al. 2008, Widiger & Simonsen 2005) and parsing functioning from symptoms and traits (Hopwood et al. 2011, Tyrer 2005). However, it is not clear whether this "hybrid" differentiates the stable and dynamic elements of PDs; such a differentiation is not made explicit in the DSM-5 proposal and represents one critical area for further research.

Future Research

More generally, although the substantial reconceptualization of PDs in DSM-5 (Skodol 2012) could potentially address many of the issues and concerns regarding the classification of PDs using current diagnostic systems, this new system requires new research on the course and outcome of these disorders. It might be hypothesized that the severity of core personality pathology fluctuates over time, perhaps related to situational or contextual factors, while the basic personality trait structure for any given individual may be more stable. From this perspective, PD features might reflect a combination of more dynamic and changing symptoms associated with core self and

interpersonal features—representing the appreciable "remission" rates seen in multiple studies of PD—and more static or slowly evolving personality trait constellations—accounting for the considerable trait stability seen in individuals diagnosed with PDs. However, future work on the etiological origins of these components of a PD system, the nature of the connections between these components, and the mechanisms by which these connections have their influence, will be critical.

The proposal also does not address a number of lingering issues with respect to the stability of PDs. There is a particular need for research on the developmental aspects of PD (Shiner 2009, Tackett et al. 2009). Childhood and adolescent experiences can have appreciable impact on the development and course of PD, and future research needs to examine personality pathology at younger ages in order to study the genesis of adult problems, perhaps developing better tools and interpretive guidelines for the detection of early personality problems. As shown in the CIC, personality pathology at a young age tends to be less extreme than during adulthood, and so this represents a crucial age at which prevention and early intervention efforts might be directed. Likewise, very little is known about the development of PD in later life, which is another critical area for future research.

Robust evidence for the psychometric superiority of dimensional scales over categorical diagnosis has raised issues that also deserve future research attention, since many important clinical decisions are binary (e.g., hospitalize or not, medicate or not, refer or not). The advantage of universal cut scores, such as those provided for categorical PD diagnosis in the DSM-IV, is that even though the cut scores may be arbitrary, they are also standardized—everyone uses the same ones. This standardization facilitates reliable professional communication. The recognition that these cut scores are unlikely to reflect any natural point of differentiation suggests the need to explore alternatives that are both faithful to the dimensional nature of PD constructs and sensitive to the demands of clinical practice. One possibility is to render a diagnosis based on functional severity, as proposed for the DSM-5. It is not clear that the answer to such debates can be completely empirical, as there does seem to be a natural tension between the underlying nature of PD constructs and the practical need to communicate efficiently about them.

Future research needs to address unanswered questions about the conflation of methods and constructs and their relative impact upon stability estimates. Although diagnostic interviews are available for traits and questionnaires are available for PDs, these instruments are used relatively infrequently and are generally regarded skeptically within their home disciplines. Specifically, psychometrically oriented personality psychologists may question the use of interviews on the basis of the additional error variance that may be associated with interviewers, and clinically oriented psychiatrists may question the ability of people to report accurately on their personality features.[3] Promising alternatives to these approaches, such as informant or performance assessments, are also infrequently used, perhaps in part due to practical difficulties associated with obtaining those kinds of data. Overall, a more comprehensive approach to assessing personality and psychopathology is needed in future studies on PD stability.

Future research should also incorporate more sophisticated sampling and analytic methods than have been used in the past. This might include sampling strategies that permit testing etiological hypotheses, such as behavior genetic designs; using planned missingness to permit a broader assessment without unduly burdening respondents; using online and other more convenient assessment approaches to maximize sample sizes; and sampling behavior at different intervals to

[3]We note, however, that from some psychometric perspectives, item responses are generally regarded as behavior samples and thus "accurate" reporting is not considered critical for valid questionnaire assessments (Meehl 1945).

evaluate issues such as dependability and substantive changes in personality processes more adequately than previously has been done. Such research should define the constructs of interest and the type of stability being conceptualized more carefully than has been the case in the past.

The observation that some features of PD do in fact change has led to exciting new areas of research involving the mechanisms and components of this change. For instance, Lenzenweger et al. (2012) recently showed that independent clusters of BPD characteristics change together during treatment and that changes in these clusters are predictable by different sets of variables. For instance, whereas baseline levels of neuroticism predicted changes in aggression, baseline levels of social potency predicted changes in conflict tolerance and behavioral control. This study, which builds upon basic observations about PD stability from the research reviewed above, is an example of the kind of work that will be important for building more precise and nuanced models of the nature of stability and change in PD features.

The past few decades had already seen a substantial increase in research, and the DSM-5 has provoked considerable recent interest. The CIC, LSPD, and MSAD continue to collect follow-along data. Other longitudinal studies of PD have been initiated, such as the St. Louis Personality and Aging Network study, which importantly sampled from an aging population (Oltmanns & Balsis 2011). The PD field will continue to draw upon longitudinal research on treatment, normative personality features, or other clinical conditions. Ideally, future studies would address the major methodological concerns reviewed above. Such studies would sample diagnostically and age-diverse individuals with frequent multimethod assessments of normal traits, pathological traits, symptoms, and functioning. Until such research answers lingering questions in the contemporary literature, estimates of the stability of PD will need to be tentative and qualified.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

LITERATURE CITED

Abrams RC, Bromberg CE. 2006. Personality disorders in the elderly: a flagging field of inquiry. *Int. J. Geriatr. Psychiatry* 21:1013–17

Am. Psychiatr. Assoc. 1952. *Diagnostic and Statistical Manual of Mental Disorders.* Washington, DC: Am. Psychiatr. Assoc. 132 pp.

Am. Psychiatr. Assoc. 1980. *Diagnostic and Statistical Manual of Mental Disorders.* Washington, DC: Am. Psychiatr. Assoc. 3rd ed. 494 pp.

Am. Psychiatr. Assoc. 1987. *Diagnostic and Statistical Manual of Mental Disorders.* Washington, DC: Am. Psychiatr. Assoc. 3rd ed., rev.

Am. Psychiatr. Assoc. 2000. *Diagnostic and Statistical Manual of Mental Disorders.* Washington, DC: Am. Psychiatr. Assoc. 4th ed., text rev. 943 pp.

Anusic I, Lucas RE, Donnellan MB. 2012. Dependability of personality, life satisfaction, and affect in short-term longitudinal data. *J. Personal.* 80:33–58

Arnett JJ. 2000. Emerging adulthood: a theory of development from the late teens through the twenties. *Am. Psychol.* 55:469–80

Baetens I, Claes L, Willemn L, Muehlenkamp J, Bijttebier P. 2011. The relationship between non-suicidal self-injury and temperament in male and female adolescents based on child- and parent-report. *Personal. Individ. Differ.* 50:527–30

Bateman AW, Fonagy P. 2000. Effectiveness of psychotherapeutic treatment of personality disorder. *Br. J. Psychiatry* 177:138–43

Biesanz JC, West SC, Kwok OM. 2003. Personality over time: methodological approaches to the study of short-term and long-term development and change. *J. Personal.* 71:905–41
Bender DS, Dolan RT, Skodol AE, Sanislow CA, Dyck IR, et al. 2001. Treatment utilization by patients with personality disorders. *Am. J. Psychiatry* 158:295–302
Bender DS, Skodol AE, Pagano ME, Dyck IR, Grilo CM, et al. 2006. Prospective assessment of treatment use by patients with personality disorders. *Psychiatr. Serv.* 57:254–57
Blashfield RK. 1993. Variants of categorical and dimensional models. *Psychol. Inq.* 4:95–98
Blashfield RK, Intoccia V. 2000. Growth of the literature on the topic of personality disorders. *Am. J. Psychiatry* 157:472–73
Bornovalova MA, Huibregtse BM, Hicks BM, Keyes M, McGue M, Iacono W. 2012. Tests of a direct effect of childhood abuse on adult borderline personality disorder traits: a longitudinal discordant twin design. *J. Abnorm. Psychol.* In press
Carpenter WT Jr, Gunderson JG. 1977. Five year follow-up comparison of borderline and schizophrenic patients. *Compr. Psychiatry* 18:567–71
Cattell RB, Eber HW, Tatsuoka MM. 1970. *Handbook for Sixteen Personality Factor Questionnaire* (*16PF*). Champaign, IL: Inst. Personal. Ability Testing. 390 pp.
Chmielewski M, Watson D. 2009. What is being assessed and why it matters: the impact of transient error on trait research. *J. Personal. Soc. Psychol.* 97:186–202
Clark LA. 2007. Assessment and diagnosis of personality disorder: perennial issues and emerging conceptualization. *Annu. Rev. Psychol.* 58:227–57
Clark LA, Harrison JA. 2001. Assessment instruments. In *Handbook of Personality Disorders*, ed. WJ Livesley, pp. 277–306. New York: Guilford
Cochran W. 1983. *Planning and Analysis of Observational Studies*. New York: Wiley. 145 pp.
Cohen J. 1992. A power primer. *Psychol. Bull.* 112:155–59
Cohen P. 1996. Childhood risks for young adult symptoms of personality disorder: method and substance. *Multivar. Behav. Res.* 31:121–48
Cohen P, Brown J, Smailes E. 2001. Child abuse and neglect and the development of mental disorders in the general population. *Dev. Psychopathol.* 13:981–99
Cohen P, Crawford TN, Johnson JG, Kasen S. 2005. The children in the community study of developmental course of personality disorder. *J. Personal. Disord.* 19:466–86
Costa PT Jr, McCrae RR. 1992. *Revised NEO Personality Inventory* (*NEO-PI-R*) *and NEO Five-Factor Inventory* (*NEO-FFI*) *Manual*. Odessa, FL: Psychol. Assess. Resourc. 101 pp.
Crawford TN, Cohen P, Brook JS. 2001. Dramatic-erratic personality disorder symptoms: I. Continuity from early adolescence into adulthood. *J. Personal. Disord.* 15:319–35
De Clercq B, De Fruyt F. 2007. Childhood antecedents of personality disorder. *Curr. Opin. Psychiatry* 20:57–61
De Fruyt F, Van Leeuwen K, Bagby RM, Rolland JP, Rouillon F. 2006. Assessing and interpreting personality change and continuity in patients treated for major depression. *Psychol. Assess.* 18:71–80
Distel MA, Willemsen G, Ligthart L, Derom C, Martin NG, et al. 2010. Genetic covariance structure of the four main features of borderline personality disorder. *J. Personal. Disord.* 24:427–44
Donnellan MB, Robins RW. 2009. The development of personality across the life span. In *Cambridge Handbook of Personality*, ed. PJ Corr, G Matthews, pp. 191–204. New York: Cambridge Univ. Press
Durbin CE, Klein DN. 2006. Ten-year stability of personality disorders among outpatients with mood disorders. *J. Abnorm. Psychol.* 115:75–84
Durbin CE, Schalet BD, Hayden EP, Simpson J, Jordan PL. 2009. Hypomanic personality traits: a multimethod exploration of their association with normal and abnormal dimensions of personality. *J. Res. Personal.* 43:898–905
Eaton NR, Krueger RF, South SC, Simms LJ, Clark LA. 2011. Contrasting prototypes and dimensions in the classification of personality pathology: evidence that dimensions, but not prototypes, are robust. *Psychol. Med.* 6:1151–63
Edens JF. 2006. Unresolved controversies concerning psychopathy: implications for clinical and forensic decision-making. *Prof. Psychol. Res. Pract.* 37:59–65
Edens JF, Marcus DK, Morey LC. 2009. Paranoid personality has a dimensional latent structure: taxometric analyses of community and clinical samples. *J. Abnorm. Psychol.* 118:545–53

Fleiss JL. 1981. *Statistical Methods for Rates and Proportions.* New York: Wiley. 2nd ed. 352 pp.
Green SB. 2003. A coefficient alpha for test-retest data. *Psychol. Methods* 8:88–101
Grilo CM, McGlashan TH. 1999. Stability and course of personality disorders. *Curr. Opin. Psychiatry* 12:157–62
Grilo CM, McGlashan TH, Oldham JM. 1998. Course and stability of personality disorders. *J. Pract. Psychiatry Behav. Health* 4:61–75
Grilo CM, Shea MT, Sanislow CA, Skodol AE, Gunderson JG, et al. 2004. Two-year stability and change in schizotypal, borderline, avoidant, and obsessive-compulsive personality disorders. *J. Consult. Clin. Psychol.* 72:767–75
Grinker RR, Werble B, Drye RC. 1968. *The Borderline Syndrome*. New York: Basic Books. 274 pp.
Guilford JP. 1954. *Psychometric Methods*. New York: McGraw-Hill. 597 pp.
Gunderson JG, Bender D, Sanislow C, Yen S, Rettew JB, et al. 2003. Plausibility and possible determinants of sudden "remissions" in borderline patients. *Psychiatry* 66:111–18
Gunderson JG, Shea MT, Skodol AE, McGlashan TH, Morey LC, et al. 2000. The Collaborative Longitudinal Personality Disorders: Study I. Development, aims, design, and sample characteristics. *J. Personal. Disord.* 14:300–15
Gunderson JG, Stout RL, McGlashan TH, Shea MT, Morey LC, et al. 2011. Ten-year course of borderline personality disorder: psychopathology and function from the Collaborative Longitudinal Personality Disorders Study. *Arch. Gen. Psychiatry* 68:827–37
Haslam N. 2003. Categorical versus dimensional models of mental disorder: the taxometric evidence. *Aust. N. Z. J. Psychiatry* 37:696–704
Haslam N, Holland E, Kuppens P. 2012. Categories versus dimensions in personality and psychopathology: a quantitative review of taxometric research. *Psychol. Med.* 42:903–20
Herpertz SC, Dietrich TM, Wenning B, Krings T, Erberich SG, et al. 2001. Evidence of abnormal amygdala functioning in borderline personality disorder: a functional MRI study. *Biol. Psychiatry* 50:292–98
Heumann K, Morey LC. 1990. Reliability of categorical and dimensional judgments of personality disorder. *Am. J. Psychiatry* 147:498–500
Hopwood CJ. 2011. Personality traits in the DSM-5. *J. Personal. Assess.* 93:398–405
Hopwood CJ, Malone JC, Ansell EB, Sanislow CA, Grilo CM, et al. 2011. Personality assessment in DSM-V: empirical support for rating severity, style, and traits. *J. Personal. Disord.* 25:305–20
Hopwood CJ, Morey LC, Donnellan MB, Samuel DB, Grilo CM, et al. 2012. Ten year rank-order stability of personality traits and disorders in a clinical sample. *J. Personal.* In press
Hopwood CJ, Morey LC, Edelen MO, Shea MT, Grilo CM, et al. 2008. A comparison of interview and self-report methods for the assessment of borderline personality disorder criteria. *Psychol. Assess.* 20:81–85
Hopwood CJ, Newman DA, Donnellan MB, Markowitz JC, Grilo CM, et al. 2009. The stability of personality traits in individuals with borderline personality disorder. *J. Abnorm. Psychol.* 118:806–15
Hopwood CJ, Zanarini MC. 2010. Five-factor trait instability in borderline relative to other personality disorders. *Personal. Disord.* 1:58–66
Huprich SK, Bornstein RF. 2007. An overview of issues related to categorical and dimensional models of personality disorder assessment. *J. Personal. Assess.* 89:3–15
Johnson JG, Cohen P, Kasen S, Skodol AE, Hamagami F, Brook JS. 2000. Age-related change in personality disorder trait levels between early adolescence and adulthood: a community-based longitudinal investigation. *Acta Psychiatr. Scand.* 102:265–75
Kass F, Skodol A, Charles E, Spitzer RL, Williams JB. 1985. Scaled ratings of DSM-III personality disorders. *Am. J. Psychiatry* 142:627–30
Katz MM, Klerman GL. 1979. Introduction: overview of the clinical studies program. *Am. J. Psychiatry* 136:49–56
Keller MB, Yonkers KW, Warshaw MG, Pratt LA, Gollan JK, et al. 1994. Remission and relapse in subjects with panic disorder and panic with agoraphobia: a prospective short-term naturalistic follow-up. *J. Nerv. Ment. Dis.* 182:290–96
Klonsky ED, Oltmanns TF, Turkheimer E. 2002. Informant-reports of personality disorder: relation to self-reports and future research directions. *Clin. Psychol.* 9:300–11

Krueger RF, Carlson SR. 2001. Personality disorders in children and adolescents. *Curr. Psychiatry Rep.* 3:46–51
Krueger RF, Eaton NR, Derringer J, Markon KE, Watson D, Skodol AE. 2011. Personality structure in the DSM-5: helping delineate personality disorder content and framing the metastructure. *J. Personal. Assess.* 93:325–31
Krueger RF, Markon KE. 2006. Reinterpreting comorbidity: a model-based approach to understanding and classifying psychopathology. *Annu. Rev. Clin. Psychol.* 2:111–33
Leichsenring F, Leibing E. 2003. The effectiveness of psychodynamic and cognitive behavior therapy in the treatment of personality disorders: a meta-analysis. *Am. J. Psychiatry* 160:1223–32
Lenzenweger MF. 1999. Stability and change in personality disorder features: the Longitudinal Study of Personality Disorders. *Arch. Gen. Psychiatry* 56:1009–15
Lenzenweger MF. 2006. The Longitudinal Study of Personality Disorders: history, design considerations, and initial findings. *J. Personal. Disord.* 20:645–70
Lenzenweger MF, Clarkin JF, Levy KN, Yeomans FE, Kernberg OF. 2012. Predicting domains and rates of change in borderline personality disorder. *Personal. Disord.* 3:185–95
Lenzenweger MF, Johnson MD, Willett JB. 2004. Individual growth curve analysis illuminates stability and change in personality disorder features: the Longitudinal Study of Personality Disorders. *Arch. Gen. Psychiatry* 61:1015–24
Lenzenweger MF, Willett JB. 2007. Modeling individual change in personality disorder features as a function of simultaneous individual change in personality dimensions linked to neurobehavioral systems: the Longitudinal Study of Personality Disorders. *J. Abnorm. Psychol.* 116:684–700
Livesley WJ. 2007. A framework for integrating dimensional and categorical classifications of personality disorder. *J. Personal. Disord.* 21:199–224
Loeber R, Dishion T. 1983. Early predictors of male delinquency: a review. *Psychol. Bull.* 94:68–99
Loranger AW, Lenzenweger MF, Gartner AF, Susman VL, Herzig J, et al. 1991. Trait-state artifacts and the diagnosis of personality disorders. *Arch. Gen. Psychiatry* 48:720–28
Maddocks PD. 1970. A five year follow-up of untreated psychopaths. *Br. J. Psychiatry* 116:511–15
Markon KE, Chmielewski M, Miller CJ. 2011. The reliability and validity of discrete and continuous measures of psychopathology: a quantitative review. *Psychol. Bull.* 137:856–79
Markon KE, Krueger RF, Watson D. 2005. Delineating the structure of normal and abnormal personality: an integrative hierarchical approach. *J. Personal. Soc. Psychol.* 88:139–57
McAdams DP, Bauer JJ, Sakaeda AR, Anyidoho NA, Machado MA, et al. 2006. Continuity and change in the life story: a longitudinal study of autobiographical memories in emerging adulthood. *J. Personal.* 74:1371–400
McCrae RR, Costa PT. 1994. The stability of personality: observations and evaluations. *Curr. Dir. Psychol. Sci.* 3:173–75
McGlashan TH. 1986. The Chestnut Lodge follow-up study. III. Long-term outcome of borderline personalities. *Arch. Gen. Psychiatry* 43:20–30
McGlashan TH. 1993. Implications of outcome research for the treatment of borderline personality disorder. In *BPD: Etiology and Treatment*, ed. J. Paris, pp. 235–59. Washington, DC: Am. Psychiatr. Press
McGlashan TH, Grilo CM, Sanislow CA, Ravelski E, Morey LC, et al. 2005. Two-year prevalence and stability of individual DSM-IV criteria for schizotypal, borderline, avoidant, and obsessive-compulsive personality disorders. *Am. J. Psychiatry* 162:883–89
Meehl PE. 1945. The dynamics of "structured" personality tests. *J. Clin. Psychol.* 1:296–303
Mihura JL, Meyer GJ, Dumitrascu N, Bombel G. 2012. The validity of individual Rorschach variables: systematic reviews and meta-analyses of the Comprehensive System. *Psychol. Bull.* In press
Mischel W. 1968. *Personality and Assessment*. New York: Wiley. 365 pp.
Morey LC, Gunderson JG, Quigley BD, Lyons M. 2000. Dimensions and categories: the "Big Five" factors and the DSM personality disorders. *Assessment* 7:203–16
Morey LC, Gunderson JG, Stout RL, Shea MT, Skodol AE, et al. 1999. *Stability of five-factor traits in personality disorder*. Presented at Annu. Meet. Am. Psychiatr. Assoc., Washington, DC
Morey LC, Hopwood CJ, Gunderson JG, Skodol AE, Shea MT, et al. 2007. Comparison of alternative models for personality disorders. *Psychol. Med.* 37:983–94

Morey LC, Hopwood CJ, Markowitz JC, Gunderson JG, Grilo CM, et al. 2012. Comparison of alternative models for personality disorders, II: 6-, 8-, and 10-year follow-up. *Psychol. Med.* 42:1705–13

Morey LC, Meyer JK. 2012. Course of personality disorder. In *The Oxford Handbook of Personality Disorders*, ed. TA Widiger, pp. 275–95. New York: Oxford Univ. Press

Morey LC, Shea MT, Markowitz JC, Stout RL, Hopwood CJ, et al. 2010. State effects of major depression on the assessment of personality and personality disorder. *Am. J. Psychiatry* 167:528–35

Morey LC, Skodol AE, Grilo CM, Sanislow CA, Zanarini MC, et al. 2004. Temporal coherence of criteria for four personality disorders. *J. Personal. Disord.* 18:394–98

Mullins-Sweatt SN, Lengel GJ, Grant DM. 2012. Nonsuicidal self-injury: the contribution of general personality functioning. *Personal. Ment. Health*. In press

Oltmanns TF, Balsis S. 2011. Personality pathology in later life: questions about the measurement, course, and impact of disorders. *Annu. Rev. Clin. Psychol.* 7:321–49

Paris J, Silk KR, Gunderson JG, Links PS, Zanarini MC. 2009. The case for retaining borderline personality as a psychiatric diagnosis. *Personal. Ment. Health* 3:96–100

Paris J, Zweig-Frank H. 2001. A 27-year follow-up of patients with borderline personality disorder. *Compr. Psychiatry* 42:782–87

Plakun EM, Burkhardt PE, Muller JP. 1985 Fourteen-year follow-up of borderline and schizotypal personality disorders. *Compr. Psychiatry* 26:448–55

Powers AD, Oltmanns TF. 2012. Personality disorders and physical health: a longitudinal examination of physical functioning, healthcare utilization, and health-related behaviors in middle-aged adults. *J. Personal. Disord.* 26:524–38

Rettew DC, Zanarini MC, Yen S, Grilo CM, Skodol AE, et al. 2003. Childhood antecedents of avoidant personality disorder: a retrospective study. *J. Am. Acad. Child. Adolesc. Psychiatry* 42:1122–30

Roberts BW, DelVecchio WF. 2000. The rank-order consistency of personality traits from childhood to old age: a quantitative review of longitudinal studies. *Psychol. Bull.* 126:3–25

Roberts BW, Walton KE, Viechtbauer W. 2006. Patterns of mean-level change in personality traits across the life course: a meta-analysis of longitudinal studies. *Psychol. Bull.* 132:1–25

Robins E, Gentry KA, Munoz RA, Marten S. 1977. A contrast of the three more common illnesses with the ten less common in a study and 18-month follow up of 314 psychiatric emergency room patients, III: findings at follow-up. *Arch. Gen. Psychiatry* 34:285–91

Rothschild L, Cleland C, Haslam N, Zimmerman M. 2003. A taxometric study of borderline personality disorder. *J. Abnorm. Psychol.* 112:657–66

Rottman BM, Ahn W, Sanislow CA, Kim NS. 2009. Can clinicians recognize DSM-IV personality disorders from Five-Factor model descriptions of patient cases? *Am. J. Psychiatry* 166:427–33

Russell JJ, Moskowitz DS, Zuroff DC, Sookman D, Paris J. 2007. Stability and variability of affective experience and interpersonal behavior in borderline personality disorder. *J. Abnorm. Psychol.* 116:578–88

Samuel DB, Hopwood CJ, Ansell EB, Morey LC, Sanislow CA, et al. 2011. Comparing the temporal stability of self-report and interview assessed personality disorder. *J. Abnorm. Psychol.* 120:670–80

Samuel DB, Widiger TA. 2008. A meta-analytic review of the relationships between the five-factor model and DSM-IV-TR personality disorders: a facet level analysis. *Clin. Psychol. Rev.* 28:1326–42

Sanislow CA, Little TD, Ansell EB, Grilo CM, Daversa M, et al. 2009. Ten-year stability and latent structure of the DSM-IV schizotypal, borderline, avoidant, and obsessive-compulsive personality disorders. *J. Abnorm. Psychol.* 118:507–19

Schmidt FL, Le H, Ilies R. 2003. Beyond alpha: an empirical examination of the effects of different sources of measurement error on reliability estimates of individual differences constructs. *Psychol. Methods* 8:208–24

Seivewright H, Tyrer P, Johnson T. 2002. Change in personality status in neurotic disorders. *Lancet* 359:2253–54

Shea MT, Stout RL, Gunderson JG, Morey LC, Grilo CM, et al. 2002. Short-term diagnostic stability of schizotypal, borderline, avoidant, and obsessive-compulsive personality disorders. *Am. J. Psychiatry* 159:2036–41

Shea MT, Yen S. 2003. Stability as a distinction between Axis I and Axis II disorders. *J. Personal. Disord.* 17:373–86

Shedler J, Beck A, Fonagy P, Gabbard GO, Gunderson JG, et al. 2010. Personality disorders in DSM-5. *Am. J. Psychiatry* 167:1026–28
Shiner RL. 2009. The development of personality disorders: perspectives from normal personality development in childhood and adolescence. *Dev. Psychopathol.* 21:715–34
Skodol AE. 2012. Personality disorders in DSM-5. *Annu. Rev. Clin. Psychol.* 8:317–44
Skodol AE, Clark LA, Bender DS, Krueger RF, Livesley WJ, et al. 2011. Proposed changes in personality and personality disorder assessment and diagnosis for DSM-5. Part I: description and rationale. *Personal. Disord.* 2:4–22
Skodol AE, Gunderson JG, Shea MT, McGlashan TH, Morey LC, et al. 2005a. The Collaborative Longitudinal Personality Disorders Study (CLPS): overview and implications. *J. Personal. Disord.* 19:487–504
Skodol AE, Oldham JE, Bender DS, Dyck IR, Stout RL, et al. 2005b. Dimensional representations of DSM-IV personality disorders: relationships to functional impairment. *Am. J. Psychiatry* 162:1919–25
Skodol AE, Pagano ME, Bender DS, Shea MT, Gunderson JG, et al. 2005c. Stability of functional impairment in patients with schizotypal, borderline, avoidant, or obsessive-compulsive personality disorder over two years. *Psychol. Med.* 35:443–51
Solhan MB, Trull TJ, Jahng S, Wood PK. 2009. Clinical assessment of affective instability: comparing EMA indices, questionnaire reports, and retrospective recall. *Psychol. Assess.* 21:425–36
Stevenson J, Meares R, Comerford A. 2003. Diminished impulsivity in older patients with borderline personality disorder. *Am. J. Psychiatry* 160:165–66
Stone MH, Hurt SW, Stone DK. 1987. The PI 500: long-term follow-up of borderline inpatients meeting DSM-III criteria. I: global outcome. *J. Personal. Disord.* 1:291–98
Tackett JL, Balsis SM, Oltmanns TF, Krueger RF. 2009. A unifying perspective of personality pathology across the lifespan: developmental considerations for DSM-V. *Dev. Psychopathol.* 21:687–713
Thomas KM, Yalch MM, Krueger RF, Wright AGC, Markon KE, Hopwood CJ. 2012. The convergent structure of DSM-5 personality trait faces and Five Factor Model trait domains. *Assessment.* In press
Trull TJ, Durrett CA. 2005. Categorical and dimensional models of personality disorder. *Annu. Rev. Clin. Psychol.* 1:355–80
Tyrer P. 2005. The problem of severity in the classification of personality disorder. *J. Personal. Disord.* 19:309–14
Watson D, Clark LA, Chmielewski M. 2008. Structures of personality and their relevance to psychopathology: II. Further articulation of a comprehensive unified trait structure. *J. Personal.* 76:1485–522
Westen D, Shedler J. 1999. Revising and assessing Axis II. Part 1: developing a clinically and empirically valid assessment method. *Am. J. Psychiatry* 156:258–72
Widiger TA. 2005. CIC, CLPS, and MSAD. *J. Personal. Disord.* 19:586–93
Widiger TA. 2007. Dimensional models of personality disorder. *World Psychiatry* 6:15–19
Widiger TA, Simonsen E. 2005. Alternative dimensional models of personality disorder: finding a common ground. *J. Personal. Disord.* 19:110–30
Wright AGC. 2011. Qualitative and quantitative distinctions in personality disorder. *J. Personal. Assess.* 93:370–79
Wright AGC, Pincus AL, Lenzenweger MF. 2012a. An empirical examination of distributional assumptions underlying the relationship between personality disorder symptoms and personality traits. *J. Abnorm. Psychol.* 121:699–706
Wright AGC, Pincus AL, Lenzenweger MF. 2012b. A parallel process growth curve model of avoidant personality disorder symptoms and personality traits. *Personal. Disord.* In press
Wright AGC, Pincus AL, Lenzenweger MF. 2012c. Interpersonal development, stability, and change in early adulthood. *J. Personal.* 80:1339–72
Wright AGC, Thomas KM, Hopwood CJ, Markon KE, Pincus AL, Krueger RF. 2012d. The hierarchical structure of DSM-5 pathological personality traits. *J. Abnorm. Psychol.* 121:951–57
Zanarini MC, Frankenburg FR, Hennen J, Reich B, Silk KR. 2005. Psychosocial functioning of borderline patients and Axis II comparison subjects followed prospectively for six years. *J. Personal. Disord.* 19:19–29
Zanarini MC, Frankenburg FR, Hennen J, Silk KR. 2003. The longitudinal course of borderline psychopathology: 6-year prospective follow-up of the phenomenology of borderline personality disorder. *Am. J. Psychiatry* 160:274–83

Zanarini MC, Frankenburg FR, Reich DB, Fitzmaurice G. 2010. Time to attainment of recovery from borderline personality disorder and stability of recovery: a 10-year prospective follow-up study. *Am. J. Psychiatry* 167:663–67

Zanarini MC, Frankenburg FR, Reich DB, Fitzmaurice G. 2012. Attainment and stability of sustained symptomatic remission and recovery among patients with borderline personality disorder and Axis II comparison subjects: a 16-year prospective follow-up study. *Am. J. Psychiatry* 169:476–83

Zanarini MC, Frankenburg FR, Reich DB, Silk KR, Hudson JI, McSweeney LB. 2007. The subsyndromal phenomenology of borderline personality disorder: a 10-year follow-up study. *Am. J. Psychiatry* 164:929–35

Zanarini MC, Skodol AE, Bender D, Dolan RT, Sanislow CA, et al. 2000. The Collaborative Longitudinal Personality Disorders Study: reliability of Axis I and II diagnoses. *J. Personal. Disord.* 14:291–99

Zimmerman M. 1994. Diagnosing personality disorders: a review of issues and research methods. *Arch. Gen. Psychiatry* 51:225–45

Zimmerman M, Chelminski I, Young D, Dalrymple K, Martinez J. 2012. Impact of deleting 5 DSM-IV personality disorders on prevalence, comorbidity, and the association between personality disorder pathology and psychosocial morbidity. *J. Clin. Psychiatry* 73:202–7

Annual Review of
Clinical Psychology
Volume 9, 2013

Contents

Indexes

Errata

An online log of corrections to *Annual Review of Clinical Psychology* articles may be found at http://clinpsy.annualreviews.org