Journal of Personality 81:3, June 2013

DOI: 10.1111/j.1467-6494.2012.00801.x

Ten-Year Rank-Order Stability of Personality Traits and Disorders in a Clinical Sample

Christopher J. Hopwood,[1] Leslie C. Morey,[2] M. Brent Donnellan,[1] Douglas B. Samuel,[3] Carlos M. Grilo,[4] Thomas H. McGlashan,[4] M. Tracie Shea,[5,6] Mary C. Zanarini,[7] John G. Gunderson,[7] and Andrew E. Skodol[8]

[1]Michigan State University
[2]Texas A&M University
[3]Purdue University
[4]Yale School of Medicine
[5]Veteran's Affairs Medical Center, Providence, RI
[6]Warren Alpert Medical School of Brown University
[7]McLean Hospital and Harvard Medical School
[8]University of Arizona School of Medicine and the Sunbelt Collaborative

Abstract

Objective: This study compares the 10-year retest stability of normal traits, pathological traits, and personality disorder dimensions in a clinical sample.

Method: Ten-year rank-order stability estimates for the Revised NEO Personality Inventory, Schedule for Nonadaptive and Adaptive Personality, and Diagnostic Interview for DSM-IV Personality Disorders were evaluated before and after correcting for test-retest dependability and internal consistency in a clinical sample ($N = 266$).

Results: Dependability-corrected stability estimates were generally in the range of .60–.90 for traits and .25–.65 for personality disorders.

Conclusions: The relatively lower stability of personality disorder symptoms may indicate important differences between pathological behaviors and relatively more stable self-attributed traits and imply that a full understanding of personality and personality pathology needs to take both traits and symptoms into account. The five-factor theory distinction between basic tendencies and characteristic adaptations provides a theoretical framework for the separation of traits and disorders in terms of stability in which traits reflect basic tendencies that are stable and pervasive across situations, whereas personality disorder symptoms reflect characteristic maladaptations that are a function of both basic tendencies and environmental dynamics.

Keywords: Personality Stability, Traits, Personality Disorders, Five Factor Theory

Stability is a defining aspect of personality traits (Asendorpf, 1992; Caspi & Roberts, 1999; Funder, 1991), and issues of personality stability are central to questions about human nature (Roberts & Caspi, 2001). Retest correlations for core personality dispositions over fairly extended intervals are rarely lower than .50, and are often in the .80s (Ardelt, 2000; Ferguson, 2010; Lucas & Donnellan, 2011; Roberts & DelVecchio, 2000; Roberts & Mroczek, 2008). Given this robust general finding, researchers have increasingly begun to focus on potential moderators of differential stability, such as the nature of the personality characteristics being examined (e.g., Ferguson, 2010). In this study, we compare 10-year stability coefficients for the scales of a normal trait questionnaire, a pathological trait questionnaire, and a personality disorder diagnostic interview in a clinical sample in order to evaluate the relative rank-order stability of normative and pathological

This study was approved by the CLPS Publication Committee and supported by NIMH grants MH 50837, 50838, 50839, 50840, 50850, 073708, and 080221. We thank C. Emily Durbin for her helpful comments on an earlier draft, and Lee Anna Clark and the University of Minnesota Press for their permission to reprint reliability estimates from the SNAP manual.

Correspondence concerning this article should be addressed to Christopher J. Hopwood, Psychology Department, Michigan State University Psychological Clinic, Psychology Building, 316 Physics-Room 107A, East Lansing, MI 48824. Email: hopwood2@msu.edu.

personality dimensions. Based on the recent suggestion that differences between personality traits and disorders can be explained by lower dependability of personality disorder assessments (Chmielewski & Watson, 2009), we also examine the effects of differential dependability across assessment methods on personality stability estimates.

The Stability of Normative and Pathological Personality

Personality stability can depend on certain attributes of the constructs being examined. For example, researchers generally find that ratings of narrow affects such as anger or sadness are less stable than ratings of broad traits such as neuroticism (Vaidya, Gray, Haig, Mroczek, & Watson, 2008). This pattern may suggest that the tendency to experience negative emotions in general is stable over time, but the experience of specific negative emotions waxes and wanes according to interactions between dispositional propensities and environmental context. Other research indicates that personality disorders (PDs) are less rank-order stable than normative traits (Durbin & Klein, 2006; Morey et al., 2007), and in fact that normative trait stability is informative about PD symptomatology (Hopwood et al., 2009; Hopwood & Zanarini, 2010a; Wright, Pincus, & Lenzenweger, 2011). PD symptoms connote functional impairments that are thought to be influenced by environmental factors (e.g., social stressors) and to respond to psychosocial interventions (e.g., psychotherapy). To the extent that contexts change differentially across individuals over time, narrower and more specific behavioral manifestations of personality features that covary within these contexts may show reduced temporal consistency relative to broad and endogenous dispositions.

This pattern parallels the separation of basic tendencies and characteristic adaptations in McCrae and Costa's (1995, 2003) five-factor theory of personality. Basic tendencies to be more or less neurotic, extraverted, open, agreeable, and conscientious are thought to mature by early adulthood, at which point they remain relatively stable in the absence of extreme environmental changes throughout adulthood. Characteristic adaptations change in response to variability in the biological or social environment. From this perspective, PD symptoms can be viewed as reflecting *characteristic maladaptations*, or maladjusted reactions to environmental contexts that are influenced by personality dispositions, previous experiences, and current contexts. These more behaviorally specific and contextual symptoms would thus be anticipated to be less stable over time than basic tendencies (Harkness & Lilienfeld, 1997; Harkness & McNulty, 2002).

However, in a recent meta-analysis, Ferguson (2010) compared stability effects from studies using broad versus narrow and normal versus pathological personality traits, and concluded that stability generally does not differ across these domains. Yet the authors of the only study used in Ferguson's meta-analysis that directly compared retest correlations for traits and PDs in the same sample (Durbin & Klein, 2006) found that PD symptom counts were less stable than normal personality traits over 10 years. This pattern was also observed in an article using data from the Collaborative Longitudinal Personality Disorders Study (CLPS) that did not contribute to Ferguson's meta-analysis.[1] Specifically, Morey et al. (2007) observed mean 2-year stability coefficients for five-factor trait domains (.74) and facets (.67) that were somewhat larger than retest correlations for PD symptom counts (.59).

Potential Effects of Reliability on Personality Stability

One potential explanation for the discrepancy between the meta-analytic conclusion drawn by Ferguson (2010) and the results of these specific studies is that unlike Ferguson, neither Morey et al. (2007) nor Durbin and Klein (2006) corrected for short-term "dependability" coefficients (i.e., retest correlations taken over a sufficiently short enough time interval that genuine trait change is deemed unlikely). Chmielewski and Watson (2009) compared short-term dependability in assessments of affects, normative traits, and schizotypal personality features and found that dependability tended to be lower for schizotypal features and affects than for normative traits. Correcting stability coefficients for dependability made the stabilities of affect measures more similar to those of trait measures. Chmielewski and Watson suggested, based on these findings, that correcting for dependability in PD ratings would close the gap between PD and trait stabilities as well. In particular, these authors argued that "transient error may have affected the findings of the Collaborative Longitudinal Study of Personality Disorders (CLPS), which documented high levels of change in PDs" (Chmielewski & Watson, 2009, p. 199). Although retest reliability data are available for the diagnostic interview in the CLPS sample (Zanarini et al., 2000), and dependability did not appreciably account for remission rates (i.e., mean-level stability) in Grilo et al. (2004), Chmielewski and Watson's (2009) hypothesis has not yet been tested with respect to rank-order stability, as we do in this study.

Another methodological issue that could affect the differential stability of personality attributes involves the properties of self-report and interview assessment methods. Many clinicians and PD researchers have raised concerns about self-report questionnaires, such as susceptibility to response bias (Huprich & Bornstein, 2007), and tend to favor diagnostic interviews for assessing personality pathology (Zimmerman et al., 1994). Conversely, most research on personality stability has been conducted using self-report questionnaires (Roberts & DelVecchio, 2000). Because personality researchers interested in developmental issues such as stability prefer questionnaires and clinicians interested in the pathological consequences of personality value interviews, the issues of

personality stability and normality versus pathology have sometimes been conflated. Indeed, Samuel et al. (2011) recently found, in the CLPS data, that PD constructs measured by self-report were more stable than PD constructs measured by diagnostic interview over 2 years, without controlling for dependability. This finding may suggest either that the interview method was less dependable than the questionnaire, or that the interview focused more on dynamic aspects of pathological personality, whereas the questionnaire tapped more stable aspects.

The Present Study

To summarize, the similarity of retest stability estimates for normative traits, pathological traits, and PDs is important both for theoretical models of personality and personality pathology and also for applied personality assessment. However, several issues about the relative stability of these different aspects of personality have interacted in previous research in a manner that has generated a lack of clarity about possible moderators, rendering the relative effects of dependability, pathology, and behavioral specificity on personality stability over extended intervals unclear. To further evaluate these issues, we assess in this study the 10-year retest stability of normative traits, pathological traits, and PDs in the CLPS sample, controlling for dependability. This study is the first evaluation of 10-year differential stability of normative personality traits, pathological personality traits, and PD symptoms in a clinical sample, and the first direct comparison of differential stability across these assessment domains that controls for dependability.

METHOD

Baseline and 10-year follow-up data from the CLPS sample were used in this study. At baseline, data were collected from 668 individuals who were selected for having borderline PD, avoidant PD, obsessive-compulsive PD, schizotypal PD, or major depressive disorder. However, the sample overall displayed extensive comorbidity (McGlashan et al., 2000) and similar levels of variability across the *DSM-IV* (American Psychiatric Association, (2000) PD symptom counts. At baseline and 10-year follow-up, participants were administered the Diagnostic Interview for DSM-IV Personality Disorders (DIPD-IV; Zanarini, Frankenburg, Sickel, & Yong, 1996) by interviewers masked to previous data; the Revised NEO Personality Inventory (NEO PI-R; Costa & McCrae, 1992), a self-report measure of the Five-Factor Model (FFM) of personality traits and facets; and the Schedule for Nonadaptive and Adaptive Personality (SNAP; Clark, 1993), a self-report measure of normative and pathological personality features. Information on the retest and internal consistency reliabilities of these measures is given below. The inter-rater reliabilities of DIPD-IV symptom counts from videotaped interviews in a baseline subsample ($N = 84$) ranged from .69 to .97 ($Mdn = .88$; Zanarini et al., 2000).

Participants from the original 668 were included in this report if they completed all three study measures at both baseline and 10-year follow-up ($N = 266$). Missing data occurred because of study attrition ($N = 237$) and because some participants who did not attrite did not complete self-report questionnaires at the follow-up ($N = 165$). We compared individuals with and without follow-up data on baseline personality variables. Although most differences were trivial, the following scales showed statistically significant ($p < .01$) but generally small (i.e., d coefficients around .20; Cohen, 1992) effect size differences between those with and without 10-year follow-up data. Individuals included in the analyses had higher scores on NEO PI-R Neuroticism ($d = .32$; each Neuroticism facet also differed between groups), DIPD-IV paranoid ($d = .31$) and schizotypal ($d = .27$), and SNAP propriety ($d = .26$), mistrust ($d = .25$), eccentric perceptions ($d = .24$), self-harm ($d = .20$), detachment ($d = .22$), and negative temperament ($d = .29$), and lower scores on NEO PI-R Extraversion ($d = .21$; particularly Positive Emotions, $d = .25$, and Gregariousness, $d = .21$, facets), openness to actions ($d = .34$), and achievement striving ($d = .21$). Overall, this pattern suggests that individuals who persisted in the study were less emotionally healthy than those who did not.

We computed retest Pearson correlations between baseline and10-year follow-up scores for each trait and PD to assess rank-order stability. Pearson correlations were selected in order to maintain comparability between the current results and those of other studies, given that the Pearson correlation is the most common metric used to represent longitudinal stability in the personality literature. Given substantial positive skew, the DIPD-IV data were square root transformed prior to analysis. Spearman's rho coefficients were computed given that these variables were skewed even after transformation. As all rho coefficients were $< |.05|$ different from Pearson's r values, we report r values to facilitate comparisons with stability coefficients from the other measures. To address potential biases associated with missing data, we recomputed all correlations for all measures using full information maximum likelihood estimation as implemented in AMOS 17.0. As no stability coefficient changed more than |.05|, we only report zero-order correlations among those with observed follow-up data.

In order to address concerns about the impact of measurement reliability on stability, we also corrected estimates using 1- to 2-week dependability coefficients for each personality variable. No retest reliability study was conducted in the CLPS sample for the NEO PI-R or SNAP. The NEO PI-R manual does not report information on its retest reliability (Costa & McCrae, 1992). Therefore, we used the retest reliability estimates from a study with a 1-week retest interval that used the NEO PI-R in a sample of 132 undergraduates (McCrae, Kurtz, Yamagata, & Terracciano, 2011). Test-retest coefficients for the SNAP trait and temperament scales are from a 1-week retest study conducted in a state hospital among 52 patients,[2] as reported in the SNAP manual (Clark, 1993). Dependability

coefficients for the DIPD-IV are from Zanarini et al. (2000), who reported correlations across a 7–10-day retest interval of PD symptom counts by different interviewers with 52 patients in the CLPS sample.

Following Ferguson (2010), we also used Cronbach's alpha internal consistency values to correct for measurement error. Although several authors (Chmielewski &Watson, 2009; McCrae et al., 2011) have argued that internal consistency corrections are theoretically and empirically inferior to retest corrections for evaluating the impact of unreliability on personality validity and stability, using internal consistency values has the advantage of allowing us to correct stability coefficients based on data provided by individuals in this sample and to directly compare the relative impacts of these two correction methods.

In addition to corrections for dependability and internal consistency, we also tested the moderating role of age and the severity of baseline functional impairment on stability using hierarchical moderated regression analyses with centered age/functioning and baseline trait values in Step 1, the interaction between age or functioning and the baseline trait in Step 2, and the 10-year value on the trait as the dependent variable. We tested age moderation because research suggests that stability varies as a function of age. In particular, young adults are often found to have somewhat less stable personalities than individuals in middle age (Ferguson, 2010; Roberts & DelVecchio, 2000), although whether age effects apply similarly to traits and PDs has not been studied extensively. Functioning may moderate stability for at least two reasons. First, some researchers have suggested that stability effects, particularly with regard to PDs, may be a function of regression to the mean reflecting distressed mood states at the baseline assessment (Chmielewski & Watson, 2009). Evidence that individuals with more severe functional impairment at baseline displayed more dramatic changes in personality features than those with less severe impairment would support this contention. Second, borderline PD, one of the PDs with more severe functional impairment (Skodol et al., 2002), has been associated with lower 6-year personality stability of FFM traits in the CLPS (Hopwood et al., 2009; $N = 432$) and lower 10-year stability of FFM traits in the McLean Study of Adult Development (Hopwood & Zanarini, 2010b; $N = 362$) samples, raising the possibility that the more functionally impaired individuals generally will exhibit lower levels of personality stability.

RESULTS

Tables 1–3 report internal consistency and dependability values, uncorrected 10-year stabilities, and 10-year stability estimates corrected for both dependability and internal consistency for the NEO PI-R, SNAP, and DIPD-IV, respectively. Overall, the dependability and internal consistency corrections yielded similar results. Given the recommendations by Chmielewski and Watson (2009) and McCrae et al. (2011), we focus here on dependability-corrected estimates. The average dependability corrected stability was .74 for the NEO PI-R domains and .70 for the facets. There was some variability across traits, with Neuroticism and Conscientiousness being somewhat less stable than the other domains. Values for the SNAP are reported in Table 2. The mean retest-corrected stability for SNAP traits was .70. Again, there was variability; for example, negative temperament and dependency had corrected retest correlations of .57, whereas this correlation was .82 for detachment and .96 for propriety. Finally, retest-corrected values for the DIPD-IV PD symptom counts are reported in Table 3. These values are substantially lower than those of the two self-report measures, with a mean of .47.

A one-way ANOVA comparing the retest-corrected stability estimates for the NEO PI-R facets, SNAP traits, and DIPD-IV PDs was used to test the hypothesis of differential stability across assessment domains directly. Prior to this analysis, we transformed the correlations to *z*-scores using the Fisher *r*-to-*z* transformation. This model was significant, $F(2, 52) = 10.67$, $p < .001$. A Duncan's multiple range post hoc analysis revealed that the interview-based PD stabilities were significantly ($p < .05$) lower than the stability estimates for the traits, with the SNAP and NEO PI-R stabilities not significantly differing. A similar model comparing alpha-corrected estimates was also statistically significant, $F(2, 52) = 13.65$, $p < .001$. However, unlike with retest-corrected estimates, post hoc tests of the alpha-corrected estimates suggested that the NEO PI-R facets were significantly more stable than the SNAP traits, which were more stable than the interview-assessed PDs. The apparent discrepancy between findings using retest- and alpha-corrected estimates across the NEO PI-R and SNAP was potentially due to retest reliabilities for the NEO PI-R (mean = .84) being generally higher than alpha reliabilities (.81) for the SNAP, whereas retest reliabilities (.83) tended to be higher than alpha reliabilities for the NEO PI-R facets (.73). In consideration of the well-known relation between alpha and scale length and the fact that mean inter-item correlations tend to be similar across the SNAP and NEO PI-R, the most likely interpretation of differences in alpha-corrected estimates is that they are a function of the SNAP's relatively longer scales. Given greater theoretical confidence in dependability (Chmielewski & Watson, 2009; McCrae et al., 2011) and the likelihood that alpha-corrected estimates were affected by scale length, we concluded that the SNAP traits and NEO facets were similarly stable, and that both are more stable than PDs.

The final set of analyses tested the potential moderating roles of age and baseline functioning on personality trait and disorder stability. There were no significant ($p < .05$) interactive effects for baseline global assessment of functioning on the stability of any trait or PD, as might be found if relatively lower stability in the DIPD-IV PDs were due to regression to the mean. There were no main effects or interactive effects of age on any of the FFM or SNAP traits. Age exerted a significant ($p < .05$) main effect, after controlling for baseline values, on obsessive-compulsive ($\beta = -.12$), narcissistic ($\beta = -.12$), and antisocial ($\beta = -.10$) PDs. These coefficients indicate that

Table 1 10-year Stability of NEO PI-R Factor and Facet Scales

		Dependability		Internal Consistency	
Trait	Stability	Reliability Estimate	Corrected Stability	Reliability Estimate	Corrected Stability
			Factors		
Neuroticism	.63	.91	.69	.91	.69
Extraversion	.71	.92	.77	.88	.81
Openness	.75	.93	.81	.90	.83
Agreeableness	.70	.92	.76	.87	.80
Conscientiousness	.63	.92	.68	.92	.68
Mean/Median	.68/.70	.82/.92	.74/.76	.90/.90	.76/.80
			Facets		
Anxiety	.50	.85	.59	.76	.66
Angry Hostility	.57	.83	.69	.82	.70
Depression	.46	.90	.51	.82	.56
Self-Consciousness	.57	.79	.72	.72	.79
Impulsiveness	.57	.77	.74	.68	.84
Vulnerability	.51	.85	.60	.79	.65
Warmth	.61	.86	.71	.79	.77
Gregariousness	.58	.89	.65	.74	.78
Assertiveness	.59	.91	.65	.79	.75
Activity	.52	.78	.67	.68	.76
Excitement Seeking	.64	.78	.82	.62	1.03
Positive Emotions	.57	.86	.66	.75	.76
Fantasy	.71	.82	.87	.82	.87
Aesthetics	.73	.91	.80	.81	.90
Feelings	.54	.82	.66	.68	.79
Actions	.60	.78	.77	.54	1.11
Ideas	.71	.87	.82	.80	.89
Values	.60	.80	.75	.64	.94
Trust	.67	.83	.81	.85	.79
Straightforwardness	.56	.86	.65	.70	.80
Altruism	.63	.75	.84	.74	.85
Compliance	.63	.83	.76	.74	.85
Modesty	.50	.86	.58	.71	.70
Tender-Mindedness	.56	.70	.80	.61	.92
Competence	.53	.80	.66	.71	.75
Order	.56	.90	.62	.70	.80
Dutifulness	.56	.75	.75	.61	.92
Ach. Striving	.61	.88	.69	.78	.78
Self-Discipline	.48	.83	.58	.82	.59
Deliberation	.55	.77	.71	.78	.71
Mean/median	.58/.57	.83/.83	.70/.70	.73/.74	.80/.79

older people had lower scores at follow-up on these three disorders, given their baseline scores, than did younger people. More pertinent to the focus of this article, age moderated the stability of four disorders: dependent ($\beta = .13$), histrionic ($\beta = -.15$), narcissistic ($\beta = -.17$), and antisocial ($\beta = -.11$). These values indicate that histrionic, narcissistic, and antisocial symptoms are somewhat more stable among younger individuals, whereas dependent symptoms are somewhat more stable among relatively older individuals.

DISCUSSION

This study compared the rank-order stability of personality features, corrected for retest dependability and internal consistency, from three prominent personality models in a clinical sample over a 10-year follow-up interval. The overall findings are presented in Figure 1 and can be summarized as follows: (a) self-attributed traits are more stable than interview-assessed personality disorders, (b) this difference holds after correcting for dependability and internal consistency estimates, and (c) dependability and internal consistency corrections generally yielded similar results.

Psychometric Explanations of Variable Stability Estimates Across Traits and Disorders

What explains the lower stability of PDs relative to traits? Although it has been suggested that CLPS findings of unex-

Table 2 10-Year Stability of SNAP Trait and Temperament Scales

		Dependability		Internal Consistency	
Trait	Stability	Reliability Estimate	Corrected Stability	Reliability Estimate	Corrected Stability
Negative temperament	.49	.86	.57	.90	.54
Mistrust	.66	.81	.81	.89	.74
Manipulativeness	.52	.77	.68	.79	.66
Aggression	.57	.89	.64	.88	.65
Self-harm	.57	.87	.66	.87	.66
Eccentric perceptions	.52	.79	.66	.85	.61
Dependency	.43	.75	.57	.82	.52
Positive temperament	.61	.89	.69	.89	.69
Exhibitionism	.61	.87	.70	.78	.78
Entitlement	.53	.80	.66	.83	.64
Detachment	.63	.77	.82	.85	.74
Disinhibition	.55	.76	.72	.80	.69
Impulsivity	.56	.74	.76	.77	.73
Propriety	.65	.68	.96	.82	.79
Workaholism	.53	.83	.64	.85	.62
Mean/median	.56/.56	.81/.80	.70/.68	.84/.85	.67/.66

Table 3 10-Year Stability of DIPD-IV Personality Disorder Symptom Counts

		Dependability		Internal Consistency	
Disorder	Stability	Reliability Estimate	Corrected Stability	Reliability Estimate	Corrected Stability
Paranoid	.39	.71	.55	.77	.51
Schizoid	.36	.52	.69	.73	.49
Schizotypal	.42	.65	.65	.75	.56
Borderline	.36	.84	.43	.78	.46
Histrionic	.15	.65	.23	.86	.17
Narcissistic	.24	.82	.29	.78	.31
Antisocial	.60	.92	.65	.73	.82
Avoidant	.42	.83	.51	.81	.52
Dependent	.23	.78	.29	.81	.28
Obsessive-compulsive	.31	.82	.38	.63	.49
Mean/median	.35/.36	.75/.80	.47/.47	.77/.78	.46/.49

pectedly low PD stability (e.g., Grilo et al., 2004; Shea et al., 2002) were due to PD assessment unreliability (e.g., Chmielewski & Watson, 2009), our analyses here revealed that PD stability remained lower than trait stability even after correcting for both retest reliability and internal consistency. This finding is consistent with the hypothesis that PD symptoms, as evaluated by interviewers, reflect characteristic maladaptations that are predisposed by traits but are also more behaviorally specific than traits. This hypothesis would suggest that symptoms would tend to vary more over time than traits as a function of environmental dynamics. The observation from our previous work (Morey et al., 2007, 2012) that PDs and traits demonstrate similar levels of criterion-related validity using methodologically balanced (i.e., including both interview-based and self-reported criterion variables) outcomes further supports this view and counters suggestions that PD stability findings in the CLPS were influenced by measurement unreliability, as psychometrically inferior scales should generally perform worse than superior scales in a predictive context.

A related concern involves the possibility that the CLPS assessment of PDs by diagnostic interview included false positives due to the ostensible effects of distressed mood on personality assessment (Chmielewski & Watson, 2009). This could lead to differential regression to the mean across true and false positive cases such that true cases may experience lower remission rates than false cases, and this in turn could potentially affect rank-order stability in addition to mean-level stability. However, previous results from the CLPS data suggest that NEO PI-R traits and DIPD-IV PDs are not differentially affected by baseline mood (Morey et al., 2010). The results from the current study also allay this concern, as the PDs for which participants were selected into the CLPS study (borderline, avoidant, obsessive-compulsive, and schizotypal) had similar stabilities (average dependability-corrected stabil-

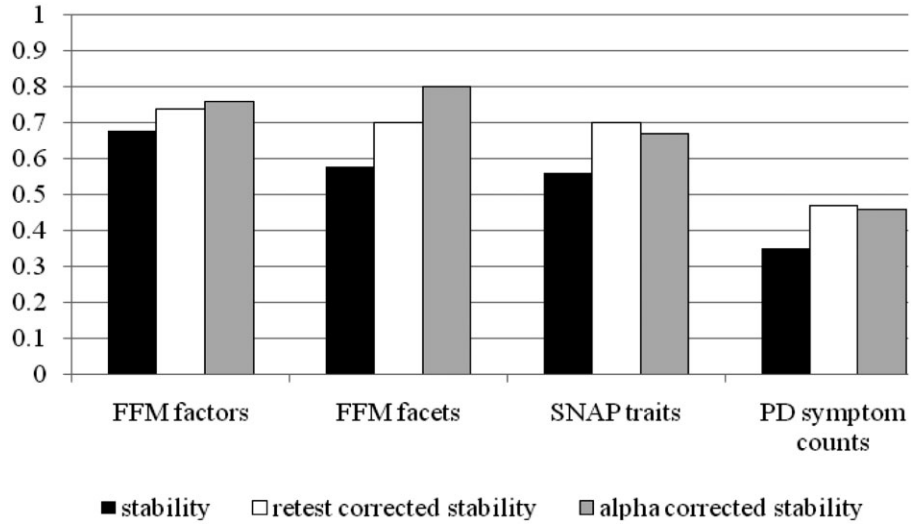

Figure 1 Mean 10-year rank-order stability values for personality traits and disorders.

ity = .49) as the PDs that were not selection criteria (.45). If regression to the mean were operative, it would be anticipated to impact selected PDs relatively more than nonselected PDs. That baseline functioning did not moderate stability for any trait or PD further discredits the hypothesis that the relatively lower rank-order instability in PDs observed in the CLPS was due to measurement invalidity.

Substantive Explanations of Varying Stability Estimates Across Traits and Disorders

There are several substantive reasons to expect that self-attributed traits might yield more stable estimates over extended periods than PD criteria. One set of explanations involves the different content of the *DSM* and trait models of personality and personality pathology. Stability differences may have occurred because the *DSM-IV* symptoms reflect a blend of pathological traits and more state-like behaviors (e.g., McGlashan et al., 2005), whereas the SNAP and FFM models more purely reflect traits and are less saturated with behaviorally specific content. Indeed, the SNAP PD scales are more stable, at least over 2 years, than the DIPD-IV PD symptom counts (Samuel et al., 2011).

PD interviews such as the DIPD-IV require clinicians to ask individuals for specific behavioral examples of their symptoms, the clinical significance of these symptoms, and how they play out in day-to-day life. Although respondents are asked to describe behavior in general and over the course of the previous 2 years, this frame of reference could magnify the importance of recent and more easily remembered experiences, and thus lead respondents and interviewers to emphasize more contextualized elements of behavior. Conversely, SNAP items are generally context free, such as "I like showing off" and "Lying is easy for me to do."[3] The nature of such questions perhaps prompts reports of more global evaluations of stable personality traits that could lead to dysfunction rather than pulling for examples of pathological behavior in specific, and possibly more recent or localized, contexts.

Interviews and questionnaires may also be differentially useful for assessing various aspects of personality and related pathology content. For example, in a previous CLPS study (Hopwood et al., 2008), our group found not only that the DIPD-IV and a self-report assessment (Personality Diagnostic Questionnaire-4) for borderline PD that was matched in terms of item content were similarly valid for predicting functional outcomes, but also that each had unique strengths. Specifically, the interview seemed somewhat more valid for more observable behavior (e.g., impulsive or self-harming behavior), but the self-report measure seemed more valid for more inferential symptoms (e.g., emptiness, identity problems). Interestingly, McGlashan et al. (2005) reported that impulsive behaviors were among the most stable interview criteria and identity problems were among the least stable. Thus, it is possible that interviewers provide less reliable and valid ratings of some aspects of PD (e.g., those that require more inference) than others.

It is also possible that different assessment methods are more amenable to the assessment of more or less stable personality features, independent of their content. For example, interviewers might be inclined to attend more closely to contextual factors that influence symptomatology, despite instructions to consider enduring aspects of personality. Clinical interviewers are trained to solicit examples of a person's behavior in addition to attributions about general behavioral tendencies, and to make judgments about whether those examples constitute significant symptomatology. Conversely, self-report questionnaires allow respondents to rate themselves on general tendencies. To the degree that it may be identity-

reaffirming to see one's personality as basically stable, self-report methods could also contribute to a level of stability over time that overestimates objective trait stability, given that self-reports are not direct measures of personality traits per se but they rather measure how people see themselves (McCrae & Costa, 1982).

Unlike self-reports, interviews require the perceptions of a clinician to filter the report of the interviewee. Because in this study different clinicians interviewed participants at baseline and follow-up, the clinicians may have noticed or rated different aspects of personality dysfunction. Thus, it is possible that variance across raters, perhaps interacting with the duration of time between interviews or interviewee state effects, lowered stability estimates somewhat. However, even if such factors were operating, these influences were not sufficiently strong to affect dependability, internal consistency, or validity coefficients (Morey et al., 2007, 2012) relative to self-report questionnaires.

LIMITATIONS AND FUTURE DIRECTIONS

This study was limited in a number of ways that suggest the need for further research. The main methodological limitation involved the high rate of unavailable data at the 10-year follow-up. Although maximum likelihood estimated stability values suggested the potentially biasing impact of missing data was limited, it is possible that the results would not generalize to individuals whose data were unavailable at the follow-up, and thus not generalize to other samples as well. It is also noteworthy that persisters showed more severe psychopathology, a finding that is worthy of more focused consideration. Despite this limitation, the CLPS data are unique in providing assessments of three measures of personality and PD assessed over 10 years and thus informative for debates about personality stability despite this limitation. Moreover, Roberts and DelVecchio (2000) reported that "across hundreds of longitudinal studies, we did not find that attrition distorted the resulting trait consistency" (p. 19).

A more methodologically balanced study could test the possibility that the observed effects were driven by measurement method. For instance, interviews might lead to lower stability estimates of personality than self-report questionnaires in general. It would be ideal to have interviews and questionnaires that were matched precisely on content. For example, it would have been useful to have interviews that assessed the content of the SNAP and NEO PI-R, and a self-report measure collected longitudinally that assessed the content of the DIPD-IV. It would also be informative for future researchers to evaluate personality stability using other methods, such as informant reports (Klonsky & Oltmanns, 2002), performance-based measures (Meyer & Viglione, 2008), or laboratory approaches (Durbin, Schalet, Hayden, Simpson, & Jordan, 2009).

Finally, more effective methods are needed for conceptualizing how measurement problems affect stability estimates. While correcting for short-term dependability is a useful way forward, it is limited in a number of ways. One example is that such corrections give less reliable measures an advantage in terms of stability estimates, in that dividing by smaller numbers leads to larger corrections. Retest dependability also assumes that the temporal difference between dependability and stability is equivalent across personality constructs, even though some personality features that can be assessed reliably (e.g., negative affectivity) are less stable than others (e.g., positive affectivity). Some PD symptoms, such as cutting behavior in borderline personality disorder, may be quite transient, and in this way be unlike some highly stable normative traits. Retest estimates were particularly limited in this study because they came from different kinds of samples across instruments. Internal consistency estimates performed similarly to dependabilities and are problematically associated with scale length, rendering internal consistency values less than optimal as a correction for reliability corrections. Intensive short-term longitudinal studies and latent variable modeling can be used to provide estimates of dependability in the presence of measurement error and help to empirically address concerns about the appropriate interval length (e.g., Anusic, Lucas, & Donnellan, 2012). The implications of this research for controlling measurement unreliability should certainly inform future investigations of personality stability.

CONCLUSION

In summary, this study showed that personality traits were substantially more consistent than PDs in a clinical sample followed over 10 years, both before and after correcting for retest dependability and internal consistency values. These results suggest that there is a continuum of stability in normal and pathological personality features, with PD symptoms reflecting relatively more contextualized and dynamic characteristic adaptations and self-attributed traits, whether normal or pathological, reflecting more stable basic tendencies. Personality traits are most useful for predicting aggregated behaviors, whereas symptoms occur in contextualized processes. Both are important and complementary for understanding personality processes and personality pathology, and this evidence, coupled with evidence of their incremental validity (Hopwood & Zanarini, 2010b; Morey et al., 2007, 2012), suggest the value of using multiple methods to assess both normal and pathological personality features for predicting behavior.

Notes

1. Two CLPS studies provided data that were considered in the meta-analysis, but both focused on the stabilities of PDs assessed by diagnostic interview from baseline to 2-year follow-up. Grilo et al. (2004)

reported remission and stability rates for categorical and dimensional PD diagnoses, and Shea et al. (2002) reported 1-year PD remission rates, mean symptom count changes, and rank-order PD stability.

2. Note that this study did not use the finalized version of the Positive Temperament and Disinhibition scales (L. A. Clark, personal communication, April 1, 2011).

3. These examples paraphrase actual SNAP items.

References

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text rev.). Washington, DC: Author.

Anusic, I., Lucas, R. E., & Donnellan, M. B. (2012). Dependability of personality, life satisfaction, and affect in short-term longitudinal data. *Journal of Personality*, **80**, 33–58.

Ardelt, M. (2000). Still stable after all these years? Personality stability theory revisited. *Social Psychology Quarterly*, **63**, 392–405.

Asendorpf, J. B. (1992). Beyond stability: Predicting inter-individual differences in intra-individual change. *European Journal of Personality*, **6**, 103–117.

Caspi, A., & Roberts, B. W. (1999). Personality change and continuity across the life course. In L. A. Pervin & O. P. John, *Handbook of personality theory and research* (Vol. 2, pp. 300–326). New York: Guilford Press.

Chmielewski, M., & Watson, D. (2009). What is being assessed and why it matters: The impact of transient error on trait research. *Journal of Personality and Social Psychology*, **97**, 186–202.

Clark, L. A. (1993). *Manual for the Schedule of Nonadaptive and Adaptive Personality.* Minneapolis: University of Minnesota Press.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, **112**, 155–159.

Costa, P.T., Jr., & McCrae, R. R. (1992). *NEO Personality Inventory-Revised professional manual*. Odessa, FL: Psychological Assessment Resources.

Durbin, C. E., & Klein, D. N. (2006). Ten-year stability of personality disorders among outpatients with mood disorders. *Journal of Abnormal Psychology*, **115**, 75–84.

Durbin, C. E., Schalet, B. D., Hayden, E. P., Simpson, J., & Jordan, P. L. (2009). Hypomanic personality traits: A multi-method exploration of their association with normal and abnormal dimensions of personality. *Journal of Research in Personality*, **43**, 898–905.

Ferguson, C. J. (2010). A meta-analysis of normal and disordered personality across the life span. *Journal of Personality and Social Psychology*, **98**, 659–667.

Funder, D. C. (1991). Global traits: A neo-Allportian approach to personality. *Psychological Science*, **2**, 31–39.

Grilo, C. M., Sanislow, C. A., Gunderson, J. G., Pagano, M. E., Yen, S. Y., Zanarini, M. C., et al. (2004). Two-year stability and change of schizotypal, borderline, avoidant, and obsessive-compulsive personality disorders. *Journal of Consulting and Clinical Psychology*, **72**, 767–775.

Harkness, A. R., & Lilienfeld, S. O. (1997). Individual differences science for treatment planning: Personality traits. *Psychological Assessment*, **9**, 349–360.

Harkness, A. R., & McNulty, J. L. (2002). Implications of personality individual differences science for clinical work on personality disorders. P. T. Costa, Jr., & T. A. Widiger (Eds.), *Personality disorders and the five-factor model of personality* (2nd ed.). Washington, DC: American Psychological Association.

Hopwood, C. J., Morey, L. C., Edelen, M. O., Shea, M. T., Grilo, C. M., Sanislow, C. A., et al. (2008). A comparison of interview and self-report methods for the assessment of borderline personality disorder criteria. *Psychological Assessment*, **20**, 81–85.

Hopwood, C. J., Newman, D. A., Donnellan, M. B., Markowitz, J. C., Grilo, C. M., Sanislow, C. A., et al. (2009). The stability of personality traits in individuals with borderline personality disorder. *Journal of Abnormal Psychology*, **118**, 806–815.

Hopwood, C. J., & Zanarini, M. C. (2010a). Borderline personality traits and disorder: Predicting prospective patient functioning. *Journal of Consulting and Clinical Psychology*, **78**, 585–589.

Hopwood, C. J., & Zanarini, M. C. (2010b). Five factor trait instability in borderline relative to other personality disorders. *Personality Disorders: Theory, Research, and Treatment*, **1**, 58–66.

Huprich, S. K., & Bornstein, R. F. (2007). Introduction to the special series in categorical and dimensional assessment of personality disorders: Implications from and for personality disorders. *Journal of Personality Assessment*, **89**, 3–15.

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

Lucas, R. E., & Donnellan, M. B. (2011). Personality development across the life span: Longitudinal analyses with a national sample from Germany. *Journal of Personality and Social Psychology*, **101**, 847–861.

McCrae, R. R., & Costa, P. T., Jr. (1982). Self-concept and the stability of personality: Cross-sectional comparisons of self-reports and ratings. *Journal of Personality and Social Psychology*, **43**, 1282–1292.

McCrae, R. R., & Costa, P. T., Jr. (1995). Trait explanations in personality psychology. *European Journal of Personality*, **9**, 231–252.

McCrae, R. R., & Costa, P. T., Jr. (2003). *Personality in adulthood: A five-factor theory perspective.* New York: Guilford Press.

McCrae, R. R., Kurtz, J. E., Yamagata, S., & Terracciano, A. (2011). Internal consistency, retest reliability, and their implications for personality scale validity. *Personality and Social Psychology Review*, **15**, 28–50.

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

McGlashan, T. H., Grilo, C. M., Skodol, A. E., Gunderson, J. G., Shea, M. T., Morey, L. C., et al. (2000). The Collaborative Lon-

gitudinal Personality Disorders Study: Baseline Axis I/II and II/II co-occurrence. *Acta Psychiatrica Scandinavica*, **102**, 256–264.

Meyer, G. J., & Viglione, D. J. (2008). An introduction to Rorschach assessment. In R. P. Archer & S. R. Smith (Eds.), *Personality Assessment* (pp. 281–336). New York: Routledge.

Morey, L. C., Hopwood, C. J., Gunderson, J. G., Skodol, A. E., Shea, M. T., Yen, S., et al. (2007). Comparison of alternative models for personality disorders. *Psychological Medicine*, **37**, 983–994.

Morey, L. C., Hopwood, C. J., Markowitz, J. C., Gunderson, J. G., Grilo, C. M., McGlashan, T. H., et al. (2012). Comparison of alternative models for personality disorders, II: 6-, 8- and 10-year follow-up. *Psychological Medicine*, **42**, 1705–1713.

Morey, L. C., Shea, M. T., Markowitz, J. C., Stout, R. L., Hopwood, C. J., Gunderson, J. G., et al. (2010). State effects of major depression on the assessment of personality and personality disorder. *American Journal of Psychiatry*, **167**, 528–535.

Roberts, B. W., & Caspi, A. (2001). Personality development and the person-situation debate: It's déjà vu all over again. *Psychological Inquiry*, **12**, 104–109.

Roberts, B. W., & DelVecchio, W. F. (2000). The rank-order consistency of personality from childhood to old age: A quantitative review of longitudinal studies. *Psychological Bulletin*, **126**, 3–25.

Roberts, B. W., & Mroczek, D. K. (2008). Personality trait stability and change. *Current Directions in Psychological Science*, **17**, 31–35.

Samuel, D. B., Hopwood, C. J., Ansell, E. B., Morey, L. C., Sanislow, C. A., Markowitz, J. C., et al. (2011). Comparing the temporal stability of self-report and interview assessed personality disorder. *Journal of Abnormal Psychology*, **120**, 670–680.

Shea, T. M., Stout, R. S., Gunderson, J., Morey, L. C., Grilo, C. M., McGlashan, T., et al. (2002). Short-term diagnostic stability of schizotypal, borderline, avoidant, and obsessive-compulsive personality disorders. *American Journal of Psychiatry*, **159**, 2036–2041.

Skodol, A. E., Gunderson, J. G., McGlashan, T. H., Dyck, I. R., Stout, R. L., Bender, D. S., et al. (2002). Functional impairment in patients with schizotypal, borderline, avoidant, or obsessive-compulsive personality disorder. *American Journal of Psychiatry*, **159**, 276–283.

Vaidya, J. G., Gray, E. K., Haig, J. R., Mroczek, D. K., & Watson, D. (2008). Differential stability and individual growth trajectories of Big Five and affective traits during young adulthood. *Journal of Personality and Social Psychology*, **76**, 267–304.

Wright, A. G. C., Pincus, A. L., & Lenzenweger, M. F. (2011). Development of personality and the remission and onset of personality pathology. *Journal of Personality and Social Psychology*, **101**, 1351–1358.

Zanarini, M. C., Skodol, A. E., Bender, D., Dolan, R., Sanislow, C., Schaefer, E., et al. (2000). The Collaborative Longitudinal Personality Disorders Study: Reliability of Axis I and II diagnoses. *Journal of Personality Disorders*, **14**, 291–299.

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

Zimmerman, M. (1994). Diagnosing personality disorders: A review of issues and research methods. *Archives of General Psychiatry*, **51**, 225–245.