Personality Disorders: Theory, Research, and Treatment
2010, Vol. 1, No. 1, 5–21

1949-2715/10/$12.00 DOI: 10.1037/a0018136

An Item Response Theory Integration of Normal and Abnormal Personality Scales

Douglas B. Samuel
Yale University School of Medicine

Leonard J. Simms
University at Buffalo, State University of New York

Lee Anna Clark
University of Iowa

W. John Livesley
University of British Columbia

Thomas A. Widiger
University of Kentucky

The *Diagnostic and Statistical Manual of Mental Disorders* (*DSM–IV–TR*) currently conceptualizes personality disorders (PDs) as categorical syndromes that are distinct from normal personality. However, an alternative dimensional viewpoint is that PDs are maladaptive expressions of general personality traits. The dimensional perspective postulates that personality pathology exists at a more extreme level of the latent trait than does general personality. This hypothesis was examined using item response theory analyses comparing scales from two personality pathology instruments—the Dimensional Assessment of Personality Pathology-Basic Questionnaire (DAPP-BQ; Livesley & Jackson, in press) and the Schedule for Nonadaptive and Adaptive Personality (SNAP; Clark, 1993; Clark, Simms, Wu, & Casillas, in press)—with scales from an instrument designed to assess normal range personality, the NEO Personality Inventory–Revised (NEO PI-R; Costa & McCrae, 1992). The results indicate that respective scales from these instruments assess shared latent constructs, with the NEO PI-R providing more information at the lower (normal) range and the DAPP-BQ and SNAP providing more information at the higher (abnormal) range. Nevertheless, the results also demonstrated substantial overlap in coverage. Implications of the findings are discussed with respect to the study and development of items that would provide specific discriminations along underlying trait continua.

Keywords: IRT, psychopathology, dimensional, personality disorder, assessment

The *Diagnostic and Statistical Manual of Mental Disorders* (*DSM–IV–TR*; American Psychiatric Association [APA], 2000) represents “the categorical perspective that Personality Disorders are qualitatively distinct clinical syndromes” (p. 689). However, an alternative perspective is that personality disorder criteria are maladaptive, extreme versions of general personality structure (Clark, 2007; Livesley, 2005; Widiger & Samuel, 2005). According to this hypothesis, items from instruments assessing the *DSM–IV* personality disorder criteria assess the same underlying constructs as general personality inventories, albeit at more extreme levels. While much research has demonstrated that instruments assessing personality pathology and those assessing normal personality traits do share common latent dimensions

Douglas B. Samuel, Department of Psychiatry, Yale University School of Medicine; Leonard J. Simms, Department of Psychology, University at Buffalo, State University of New York; Lee Anna Clark, Department of Psychology, University of Iowa; W. John Livesley, Department of Psychiatry, University of British Columbia; Thomas A. Widiger, Department of Psychology, University of Kentucky.

This research was partially supported by a fellowship (F31MH074245-02) from the National Institute of Mental Health, awarded to Douglas B. Samuel. Douglas B. Samuel is the author of the Dimensional Assessment of Personality Pathology-Basic Questionnaire. Lee Anna Clark is affiliated with the Schedule of Nonadaptive and Adaptive Personality. She has no financial interest in the product.

Correspondence concerning this article should be addressed to Douglas B. Samuel, VA Connecticut Healthcare System–Building 35, 950 Campbell Avenue (151-D), West Haven, CT 06516. E-mail: douglas.samuel@yale.edu

(Markon, Krueger, & Watson, 2005; O'Connor, 2005; Schroeder, Wormsworth, & Livesley, 1992), there has been very little research that has tested whether personality disorder instruments assess the shared traits at more extreme levels than general personality instruments. However, such a comparison is possible using item response theory (IRT).

The field of psychological assessment has been based largely in classical test theory (CTT), but significant advances in psychometrics have led to improved techniques for developing and evaluating assessment instruments. A primary example of these advances is the application of IRT (for a detailed history and description of IRT, see Embretson & Reise, 2000). IRT was first introduced to psychology by way of educational testing, as a method of developing more efficient measures of educational attainment or achievement. Only recently has it been applied to personality assessment, primarily to develop computerized adaptive testing (CAT) versions of existing measures. For example, Reise and Henson (2000) reported on a CAT version of the NEO Personality Inventory–Revised (NEO PI-R; Costa & McCrae, 1992) using a real-data simulation, and Simms and Clark (2005) developed and validated an IRT-based CAT for the Schedule for Nonadaptive and Adaptive Personality (SNAP; Clark, 1993).

Another potentially useful extension of IRT to the study of personality and personality pathology is its ability to compare the amount of information that existing instruments provide at different levels of a latent trait (Reise & Henson, 2000).[1] Items typically vary in the amount of information they provide across levels of a trait. For example, some items may provide little information at low levels of a trait (e.g., all persons within the lower range provide the same answer), but a great deal of information at higher levels (i.e., persons at the higher levels of the trait respond differentially to the item). Thus, as long as items from different measures can be shown to load on the same latent dimension, they can be compared in terms of the levels of that latent trait where they provide the greatest discrimination. It is this aspect of IRT that could be used to compare where measures of normal and abnormal personality functioning provide more or less information along an underlying latent continuum.

We are aware of only two published studies that have used IRT in this manner (Walton, Roberts, Krueger, Blonigen, & Hicks, 2008; Zickar, Russell, Smith, Bohle, & Tilley, 2002; and the latter is not directly pertinent to the current study as it addresses time-of-day preferences for working). Walton and colleagues administered the Psychopathic Personality Inventory (PPI; Lilienfeld & Andrews, 1996) and the Multidimensional Personality Questionnaire (MPQ; Tellegen, 1982). Subsets of items from each instrument that were deemed to assess the same underlying construct were subjected to IRT analysis. Information curves were calculated for each instrument and presented on a common graph. Walton and colleagues concluded from a visual inspection of these curves that, inconsistent with the hypothesis that psychopathy items were assessing extreme variants of normal personality, the instruments did not appear to be assessing different regions of the latent trait. This may be because the PPI assesses personality traits that are risk factors for psychopathy rather than psychopathic personality per se. These two studies have demonstrated a novel use of IRT techniques; however, given that the results of Walton et al. (2008) were inconsistent with hypotheses, a closer examination of normal and abnormal personality scales is warranted.

One of the most heavily researched models of general personality functioning is the Five Factor Model (FFM; McCrae & Costa, 1999), which has the five broad dimensions of extraversion (vs. introversion), agreeableness (vs. antagonism), neuroticism (vs. emotional stability), conscientiousness (vs. undependability), and openness (vs. closedness to experience). Over the past two decades, the FFM has provided a useful dimensional framework for understanding the *DSM* personality disorders, and well over 50 published studies support the link between them (Widiger & Costa, 2002). A meta-analysis of a number of these studies (Samuel & Widiger, 2008), reviews of this research (Clark, 2007; Livesley, 2001), and an interbattery factor analysis of published data

[1] In a literal sense, the phrase "IRT-based analyses" should be used (rather than simply "IRT") in this sentence and many others in this article. However, it is most common in the literature to refer to both the theory and its application simply by "IRT" and we shall follow this usage.

sets that examined relations between the FFM and the personality disorders (O'Connor, 2005) all have led to the conclusion that there are strong and robust links between the *DSM–IV* PD formulations and dimensions of normal personality. Thus, the FFM is a compelling candidate to assess general personality traits within an IRT-based comparison.

Two instruments measuring maladaptive personality traits that would lend themselves to an integrated IRT study of the common latent structure underlying normal and abnormal personality are the Dimensional Assessment of Personality Pathology–Basic Questionnaire (DAPP-BQ; Livesley & Jackson, in press) and the SNAP (Clark, 1993; Clark et al., in press). Both instruments were derived through an iterative process that included factor analyses of personality disorder symptomatology.

Exploratory factor analytic studies have demonstrated empirically that the dimensions of maladaptive personality functioning assessed by the DAPP-BQ and the SNAP are well integrated with at least four of the five domains of the FFM (Clark & Livesley, 2002). For example, Schroeder and colleagues (1992) examined the DAPP-BQ and the NEO PI (Costa & McCrae, 1985) in a sample of 300 community members, and found that four components of a five-component solution mapped cleanly onto the FFM (i.e., neuroticism, extraversion, agreeableness, and conscientiousness), whereas the "openness" domain included a considerable loading of scales assessing extraversion.

Clark and colleagues (in press) report the results of three principal factor analyses of the SNAP and various FFM measures, each of which yielded a five-factor solution that closely mirrored four of the five domains of the FFM. An openness factor emerged more strongly than in the FFM/DAPP-BQ analyses, but still was defined inconsistently across the three samples. Thus, Clark and colleagues (in press) concluded that openness was not well represented in the SNAP item pool. Clark, Livesley, Schroeder, and Irish (1996) also provided evidence for the convergence of the DAPP-BQ and the SNAP via joint exploratory factor analysis of the two instruments. Five factors were extracted, four of which corresponded well to neuroticism, extraversion, agreeableness, and conscientiousness, whereas, similar to the previous studies, openness appeared not to be well represented in either instrument.

Finally, Markon et al. (2005) conducted a series of exploratory factor analyses of a meta-analytically derived correlation matrix as well as new data sets that included the DAPP-BQ, SNAP, NEO PI-R, and other measures of normal and abnormal personality functioning. This study explored how normal and abnormal personality scales might be integrated within a common hierarchical structure. The authors concluded that their "results reinforce the position that the Big Five represent a crucial level of analysis for normal personality research and extend this position to include psychopathology research as well" (p. 154). Of specific relevance to the current study was the further empirical documentation of a common underlying trait structure among the DAPP-BQ, SNAP, and NEO PI-R scales.

In summary, extensive research supports the view that these three measures of general and maladaptive personality functioning share a common four-factor structure and therefore could be amenable to IRT analysis. These IRT analyses will provide data on whether personality pathology instruments assess the shared latent traits at more extreme levels than general personality measures. For example, to the extent that the DAPP-BQ compulsivity and SNAP workaholism scales measure maladaptive extreme variants of NEO PI-R conscientiousness, the compulsivity and workaholism items should provide more psychometric information at higher (i.e., more severe) levels of the underlying trait than the NEO PI-R items. In turn, the NEO PI-R items should provide more information at lower (i.e., less severe) levels of the trait. Comparable hypotheses can be made for other DAPP-BQ, SNAP, and NEO PI-R items with respect to the three additional broad latent factors underlying their integration identified in previous research.

Method

Samples and Participants

The data for the current study were drawn from two separate data collections. The first included 920 individuals who were administered the DAPP-BQ and the NEO PI-R as part of an adult community sample collected in British Columbia, Canada (Jang, Livesley, & Vernon, 2002). The sample was predominantly female (63%) and had a mean age of 33.6 years (*SD* = 13.8). A second sample included 680 students at the University of Kentucky with a mean age of 19.8 years (*SD* = 4.4) who completed the SNAP and the NEO PI-R to fulfill course credit. The majority was female (62%) and Caucasian (85%), with 10% African Americans and 5% other ethnic groups. Portions of this dataset have been used in previously published studies (e.g., Mullins-Sweatt, Jamerson, Samuel, Olson, & Widiger, 2006). Descriptive statistics for the scales from each sample are presented in online Appendix A.

Measures

NEO PI-R. The NEO PI-R (Costa & McCrae, 1992) is a measure of the FFM and contains 240 items that are rated on a Likert-type scale ranging from 1 (*strongly disagree*) to 5 (*strongly agree*). This instrument is comprised of five broad domain scales—neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness—each of which is assessed by six underlying facet scales. Internal consistency is high for the domains (coefficient alphas range from .86 to .95) and moderate to strong (.56 to .81) for the facet scales (Costa & McCrae, 1992). The NEO PI-R has evidenced strong temporal stability, with values ranging from .76 to .84 over a 7-year period (Costa, Herbst, McCrae, & Siegler, 2000).

DAPP-BQ. The DAPP-BQ (Livesley & Jackson, in press) contains 560 statements to which an individual responds on a 5-point Likert-type scale ranging from "*strongly disagree*" to "*strongly agree*." The DAPP-BQ includes 18 scales developed in part through factor analysis of personality disorder symptomatology; for example, affective lability, social avoidance, conduct problems, and compulsivity. These scales are internally consistent with coefficient alphas ranging from .83 to .94 and reliable over 3-week intervals with test–retest reliability ranging from .81 to .93 (Larstone, Jang, Livesley, Vernon, & Wolf, 2002).

SNAP-2. The SNAP-2 (Clark et al., in press) is a 390-item instrument that uses a True-False format. The SNAP includes 12 lower-order trait scales developed in part through factor analyses of personality disorder symptomatology, such as self-harm, aggression, manipulation, and workaholism. It also includes three "temperament" scales that assess, respectively, the core of three higher order personality domains, Negative Temperament, Positive Temperament, and Disinhibition, but these scales were not used in the current study. SNAP scales are internally consistent (median coefficient alphas ranged from .76 to .92; median = .81 in samples of patients and nonpatient college students and adults) and are stable over short (1-week; retest *r* range = .81 to .93; median = .88) and moderate (up to 4 months; retest *r* range = .76 to .89; median = .85) intervals (Clark et al., in press).

Domain Configuration

We first arranged the items from each instrument into the four domains consistent with the prior factor analytic research discussed earlier: (1) emotional instability, (2) antagonism, (3) introversion, and (4) constraint. In accordance, all items were keyed in the direction of these constructs. This grouping was done at the scale level because each of the four domains typically was comprised of all items from one NEO PI-R domain and all items from related DAPP-BQ scales. For example, one of the four domains identified by the joint factor analytic studies included the neuroticism scale from the NEO PI-R as well as the anxiety, suicidal ideation, insecure attachment, affective lability, identity disturbance, and submissiveness scales from the DAPP-BQ (Markon et al., 2005; Schroeder et al., 1992). Similarly, the items from the SNAP scales of detachment and exhibitionism were joined with NEO PI-R extraversion scale to form a group hereafter referred to as the introversion domain.

Analytic Procedures and Results

Unidimensionality Assessment

An assumption underlying IRT models is that items being analyzed form a unidimensional latent construct. This was particularly important in the current study because the items were obtained from different instruments. Because the SNAP utilizes a dichotomous, "true-false" format, the NEO PI-R items were also dichotomized so that they could be compared in a straightforward manner; specifically, responses "strongly disagree" and "disagree" were recoded as false, whereas "agree" and "strongly agree" were recoded as true. Also, because the recoding of the NEO PI-R's "neutral" response option has the potential to affect our results, we chose to recode "neutral" as "false" to provide a conservative and more stringent test of our hypotheses (i.e., recoding in this way makes the items more "difficult" in an IRT framework and thus biases them slightly toward abnormality).

Stout (1987, 1990) has argued that what is required for IRT is not the absence of any subfactors, but the presence of a single, dominant factor that is common to the items. Thus, we sought to demonstrate that the underlying traits were essentially unidimensional for the purposes of IRT, meaning that a broad, general dimension underlies all item responses.

We used the MicroFACT 2.0 (Waller, 2002) software program to compute three statistics assessing the fit of a one-factor model to the data in each of the four domains. Consistent with past studies, we first calculated the ratio of the first to second eigenvalue of the polychoric correlation matrix to assess the presence of a dominant first factor. Additionally, we calculated the goodness-of-fit index (GFI), an indicator of absolute fit for a one-factor solution. Values over .90 are considered to be evidence of good fit and those over .95 indicative of an excellent fit (Hu & Bentler, 1999). Finally, we also calculated the root-mean-square residual (*RMSR*), for which lower values indicate better model-to-data fit and those under .10 suggest essential unidimensionality. Although there is no infallible statistical indicator of latent structure, these three measures taken together are considered to provide adequate information to evaluate the assumption of essential unidimensionality (Stout, 1987).

These unidimensionality indices indicated that only one of the newly sorted domains (e.g., DAPP-BQ/NEO PI-R emotional instability) evidenced essential unidimensionality; the others were insufficient for IRT analyses. When unidimensionality was not clearly evidenced, items with low (typically ≤.50) factor loadings were removed. We examined these deleted items for possible inclusion in another domain, but none was retained. Following these deletions, the remaining items were reassessed for unidimensionality.

Table 1 presents the final unidimensionality results for each domain from the two sets of comparisons. The values in Table 1 are generally at or above the criteria indicating essential unidimensionality (Stout, 1987) and are comparable to those reported in previous IRT studies with measures of personality (Jane et al., 2007; Reise & Henson, 2000; Reise, Smith, & Furr, 2001; Simms & Clark, 2005). The ratio of the first to second eigenvalues in the current study ranged from 4.1 to 6.7, which compares well with previous reports of this statistic (Bolt, Hare, Vitale, & Newman, 2004; Cooke & Michie, 1997; Jane et al., 2007; Reise et al., 2001; Smith & Reise, 1998). We further investigated the presence of a dominant first factor by examining a scree plot for each of the domains (see online Appendix B). Together with the unidimensionality values presented in Table 1, these scree plots indicate that the newly created domains met Stout's (1990) criteria for essential unidimensionality and were amenable to IRT analysis.

Content Analysis

The process of refining these scales inherently reduced the number of items that were analyzed. While this reduction was helpful in satisfying unidimensionality, it raised questions about whether the resulting scales were sufficiently similar in content to

Table 1
Unidimensionality Assessment Values

	Ratio	GFI	RMSR
DAPP-BQ			
Emotional instability (140)	6.7	0.94	0.076
Antagonism (114)	4.7	0.89	0.091
Introversion (31)	5.6	0.96	0.089
Constraint (73)	5.0	0.93	0.083
SNAP			
Emotional instability (30)	4.4	0.93	0.111
Antagonism (51)	4.6	0.91	0.107
Introversion (26)	5.5	0.96	0.091
Constraint (37)	4.1	0.93	0.108

Note. Ratio = the ratio of the first to second eigenvalue; GFI = goodness-of-fit index; RMSR = root mean squared; DAPP-BQ = the Dimensional Assessment of Personality Pathology–Basic Questionnaire; SNAP = Schedule for Nonadaptive and Adaptive Personality. The first heading indicates the instruments being compared with the NEO Personality Inventory–Revised (NEO PI-R), and the indented heading indicates the domain being examined. The number within the parentheses is the final number of items included within the domain for that particular analysis.

the originals. To investigate this possibility, we examined the content from those items that were retained and compared it to those items that were excluded. In other words, we examined whether (a) the remaining items within the SNAP/NEO PI-R and DAPP-BQ/NEO PI-R comparisons were faithful to the original item pools and (b) the remaining items within these two comparisons were similar to each other. To do so, we first counted the number of items that were retained from each facet of the NEO PI-R and the SNAP and DAPP-BQ scales (see online Appendix C). From these counts, it appeared that certain scales and facets were more strongly represented than others. Not surprisingly, it appeared that the items from NEO PI-R facets most closely related in content to the DAPP-BQ and SNAP scales were more likely to be retained.

For example, these content analyses evidenced that the remaining emotional instability construct for the SNAP and NEO PI-R appears to be largely defined by negative mood and hopelessness and is somewhat narrower than the broader construct including anxiousness, anger, and impulsiveness that emerged from the comparison of the DAPP-BQ and NEO PI-R. For the introversion domain it appeared that the construct was quite similar across both comparisons and was characterized by social withdrawal and emotional coldness. Although there were differences in the conceptualizations of antagonism, depending largely on the content of the DAPP-BQ and SNAP, it appeared that the constructs were quite similar across the two comparisons and the remaining items had high fidelity with the original content. Finally, some dissimilarity was again noted for the constraint domain as a majority of the items from both measures were retained for the DAPP-BQ/NEO PI-R comparison, suggesting that the meaning of the combined dimension changed little from its component parts. However, a number of items were excluded across the NEO PI-R facets and SNAP scales such that the SNAP/NEO PI-R comparison was perhaps more heavily laden with impulsivity than the DAPP-BQ/NEO PI-R comparison.

IRT Analyses

We chose Samejima's (1969) graded response model to estimate the item parameters for the analyses of the DAPP-BQ and the NEO PI-R within this comparison because both instruments use a 5-point Likert scale. The graded response model is an extension of the two-parameter logistic model for polytomous items and is commonly used for IRT analyses of personality instruments with Likert-type scales (e.g., Reise & Henson, 2000). As the SNAP and NEO PI-R items were both dichotomous for this comparison (using the procedures outlined earlier), these analyses were conducted using the two-parameter logistic model. The IRT parameters for all analyses were estimated using Multilog 7.03 (Thissen, Chen, & Bock, 2003).

The primary results of interest for the current study are the item information curves (IICs), which show the amount of psychometric information that each item provides at all levels of the latent trait. An important property of item information curves is that they can be summed or averaged to provide an overall estimate of measurement precision for a complete scale across all levels of the underlying construct. Because total information curves are sensitive to scale length, simply summing the information curves for a scale would have produced results that were biased by the number of items retained for the scale. In order to place each scale and instrument on a more level playing field we chose to average the IICs to control for length. We termed these "mean information curves" (MICs). Within the DAPP-NEO and SNAP-NEO domains, separate MICs were calculated for each of the scales that comprised the domain. For example, separate MICs were calculated for the SNAP scales of detachment and exhibitionism as well as for the NEO PI-R domain of extraversion. The Multilog software provides an estimate of the psychometric information at levels of theta ranging from -3.0 to 3.0, at intervals of .02. Thus, the mean item information values were tested among scales at each interval through a series of one-way analyses of variance (ANOVAs), with Tukey post hoc contrasts. This allowed for a statistical comparison at each interval of theta to determine whether scales were providing different levels of information. Because space limitations preclude presentation of all possible MICs, we provide only a few illustrative examples. When examining these MICs we were looking for notable differences in the height of the curve at different points along the continuum. The greater the distance between the two lines the greater the difference between the amount of information that each respective scale is providing at that particular level of the trait.

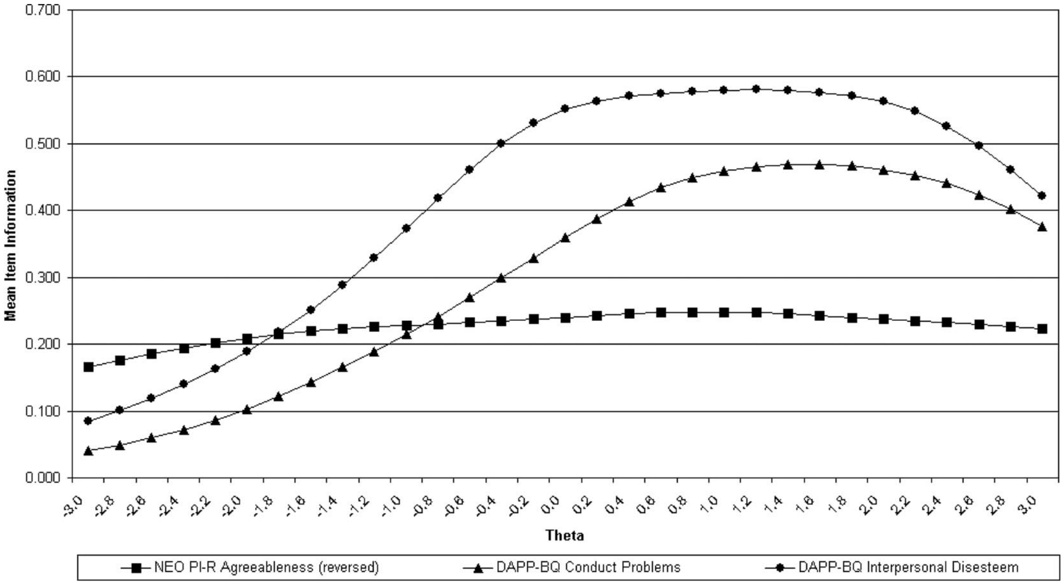

Figure 1. DAPP-BQ and NEO PI-R antagonism mean information curves.

Figure 1 presents the MICs for selected scales from the DAPP-NEO antagonism domain. In this case, the NEO agreeableness curve appeared to be relatively flat, indicating that it was providing roughly equivalent information at all levels of antagonism. By contrast, the two scales from the DAPP-BQ provided increasing psychometric information as the level of theta increased. Statistical comparisons indicated that the curve for the NEO PI-R agreeableness scale was significantly higher than both DAPP-BQ scales from thetas of −3.0 to −2.6. Similarly, the DAPP-BQ conduct problems scale provided significantly more information than either DAPP-BQ interpersonal disesteem or NEO PI-R agreeableness at levels of theta from −1.2 to 0. Finally, the curves for both of the DAPP-BQ scales were significantly higher than the NEO PI-R agreeableness curve at levels of theta above 0.4, indicating that these DAPP-BQ items provided more information at the highest levels of the latent trait.

Figure 2 presents the MICs for the scales comprising the SNAP-NEO introversion domain. A visual inspection of this figure indicated that the curves for all three of these scales peak at roughly the same level of the latent trait. However, while their location along theta was comparable, there were differences between the scales in terms of the amount of information provided. This was particularly evident at levels of theta ranging from 0.4 to 1.4, where the SNAP detachment scale provided significantly more information than the NEO PI-R extraversion scale, whereas the NEO PI-R scale provided more information than the SNAP exhibitionism scale. Thus, the findings from this particular analysis did not suggest that the SNAP and NEO PI-R provided information at different levels of the latent trait. Rather, they indicated that the SNAP detachment scale provided more fidelity in assessing introversion within a selected range of theta.

Figure 3 presents selected MICs from the DAPP-NEO emotional instability domain. A visual inspection of these curves suggested that DAPP-BQ affective lability and NEO PI-R neuroticism covered the latent trait of emotional instability in very similar ways. In fact, both of these scales evidenced a moderate assessment of the latent trait across all levels of theta and were not significantly different from one another. However, the curve for the suicidal ideation scale from the DAPP-BQ provided virtually no psychometric information at the lower levels of theta before spiking upward to provide a great deal of information at the highest levels of theta. The results of a one-way ANOVA

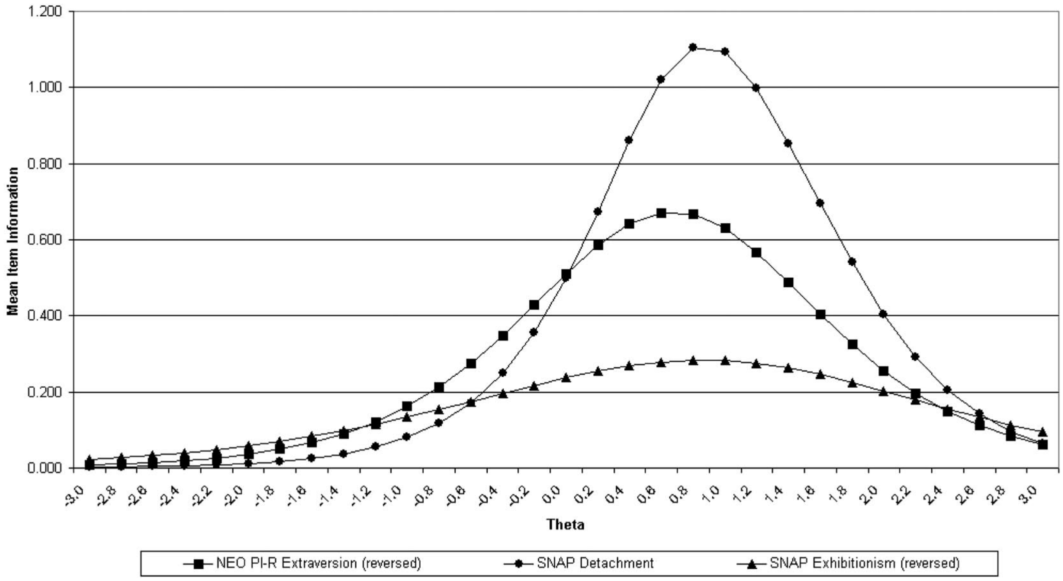

Figure 2. SNAP and NEO PI-R introversion mean information curves.

indicated that the DAPP-BQ suicidal ideation curve was significantly lower than both DAPP-BQ affective lability and NEO PI-R neuroticism at the levels of theta from −3.0 to 0.4. Conversely, the suicidal ideation curve was higher than both of these scales at all levels of theta above 1.4. Given this curve, it appeared that the suicidal ideation scale provided discrimination only among individuals who were quite extreme on the trait of emotional instability.

In lieu of providing MICs for every scale within each set of comparisons, Tables 2 and 3 present the mean alpha and beta parameters

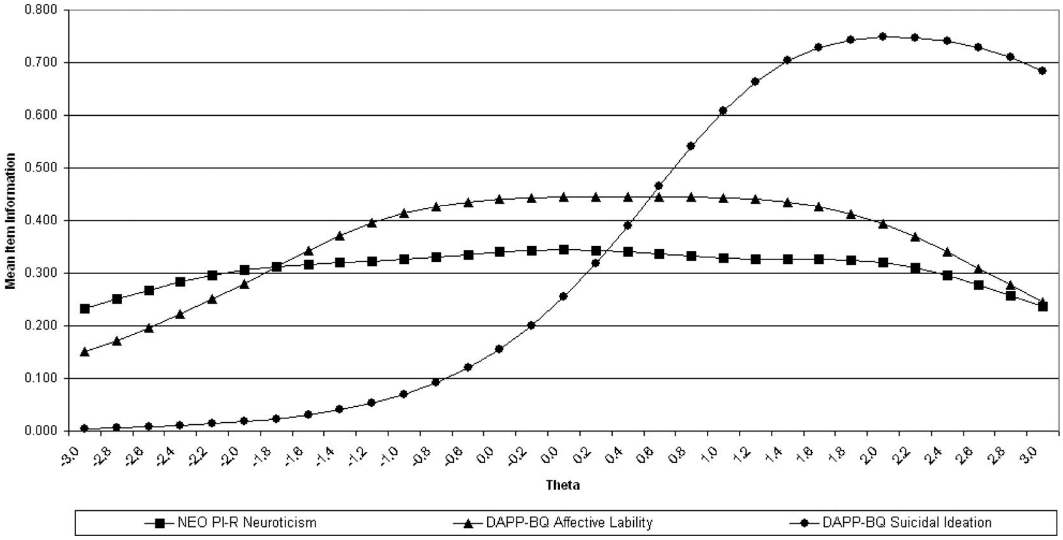

Figure 3. DAPP-BQ and NEO PI-R emotional instability mean information curves.

Table 2

Comparisons of Beta Parameters for NEO Personality Inventory–Revised (NEO PI-R) and Dimensional Assessment of Personality Pathology–Basic Questionnaire (DAPP-BQ) Scales

	Alpha		b1		b2		b3		b4	
	M	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
DAPP-NEO Emotional Instability										
1. NEO Neuroticism (48/48)	1.02	.11	−3.80	.44	−0.76	.18	0.54	.20	3.50	.56
2. DAPP Insecure Attachment (16/16)	0.97	.10	−1.22	.16	0.40	.16	1.85	.30	3.55	.54
3. DAPP Anxiety (16/16)	1.57	.13	−1.34	.10	−0.25	.08	0.76	.12	1.96	.21
4. DAPP Affective Lability (16/16)	1.14	.11	−1.82	.17	−0.42	.11	0.91	.17	2.38	.33
5. DAPP Identity Disturbance (16/16)	1.47	.12	−0.89	.08	0.30	.10	1.45	.18	2.64	.32
6. DAPP Submissiveness (16/16)	0.90	.09	−2.33	.26	−0.52	.14	1.46	.26	3.81	.58
7. DAPP Suicidal Ideation (12/12)	1.51	.22	1.39	.23	1.91	.30	2.53	.39	3.32	.58
Significant differences within column	3 > 1,2,4,6 5,7 > 1,2,6		7 > 1,2,3,4,5,6 2,3,4,5,6,7 > 1		7 > 1,2,3,4,5,6 2,5,7 > 1		7 > 1,3,4,5,6 2 > 1,3,4 2,5,6,7 > 1		3 > 1,2	
DAPP-NEO Introversion										
1. NEO Extraversion (13/48)	1.30	.11	−1.59	.13	0.62	.11	1.75	.19	3.59	.44
2. DAPP Intimacy Problems (3/16)	1.12	.11	−0.45	.09	0.91	.14	2.34	.27	3.68	.46
3. DAPP Restricted Expression (5/16)	1.21	.11	−1.19	.11	0.12	.10	1.39	.16	2.76	.30
4. DAPP Social Avoidance (10/16)	2.02	.14	−0.73	.06	0.13	.06	1.07	.18	2.05	.18
Significant differences within column	4 > 1,2,3		2,4 > 1		*ns*		2 > 4		1,2 > 4	
DAPP-NEO Antagonism										
1. NEO Agreeableness (38/48)	0.89	.11	−2.69	.34	0.56	.27	2.26	.41	5.07	.88
2. DAPP Narcissism (15/16)	1.12	.11	−1.85	.16	−0.66	.11	0.67	.17	2.30	.34
3. DAPP Suspiciousness (14/14)	1.45	.13	−0.93	.17	0.26	.13	1.40	.20	2.71	.40
4. DAPP Conduct Problems (16/16)	1.19	.15	0.18	.14	1.02	.22	1.85	.31	2.84	.46
5. DAPP Rejection (15/16)	1.01	.11	2.40	.28	−0.81	.15	0.77	.19	2.78	.45
6. DAPP Interpersonal Disesteem (16/16)	1.33	.13	−0.59	.10	0.66	.15	1.77	.26	2.99	.44
Significant differences within column	4,6 > 1 3 > 1,5		4,6 > 1,2,5 3 > 1,5		1,4,6 > 2,5		1 > 2,5		1 > 2,3,4,5	
DAPP-NEO Constraint										
1. NEO Conscientiousness (45/48)	1.13	.11	−3.95	.59	−1.27	.22	0.02	.15	2.68	.30
2. DAPP Compulsivity (13/16)	1.19	.11	−2.92	.40	−1.45	.20	0.29	.13	2.09	.24
3. DAPP Passive Aggressive (15/16)	1.38	.12	−2.62	.30	−1.31	.16	−0.17	.09	1.03	.11
Significant differences within column	*ns*		2,3 > 1		*ns*		*ns*		1 > 3	

Note. *ns* = not signficant. Significance value set at $p < .05$; b1, b2, b3, and b4 = 1st, 2nd, 3rd, and 4th beta parameters, respectively. The proportion in the parentheses following each scale name indicates the number of items retained (numerator) from the number of items in the original scale (denominator).

Table 3
Comparisons of Beta Parameters for NEO Personality Inventory–Revised (NEO PI-R) and Schedule for Nonadaptive and Adaptive Personality (SNAP) Scales

	Alpha		Beta	
	M	*SE*	*M*	*SE*
SNAP-NEO Emotional Instability				
1. NEO Neuroticism (12/48)	1.50	.22	0.84	.14
2. SNAP Self-Harm (12/16)	1.70	.28	1.41	.18
3. SNAP Dependency (6/18)	1.04	.19	1.15	.23
Significant differences within column	1,2 > 3		2 > 1	
SNAP-NEO Introversion				
1. NEO Extraversion (12/48)	1.68	.23	0.68	.10
2. SNAP Detachment (7/18)	2.17	.29	0.92	.10
3. SNAP Exhibitionism (7/16)	1.10	.18	0.74	.16
Significant differences within column	2 > 1 > 3		*ns*	
SNAP-NEO Antagonism				
1. NEO Agreeableness (18/48)	1.26	.20	0.44	.15
2. SNAP Manipulativeness (11/20)	1.40	.23	0.78	.15
3. SNAP Entitlement (3/16)	1.05	.19	1.05	.21
4. SNAP Mistrust (4/19)	1.08	.18	0.72	.17
5. SNAP Aggression (15/20)	1.68	.28	1.00	.15
Significant differences within column	5 > 1,4		5 > 1	
SNAP-NEO Constraint				
1. NEO Conscientiousness (24/48)	1.68	.47	−1.05	.43
2. SNAP Impulsivity (9/19)	1.19	.25	−0.64	.29
3. SNAP Workaholism (4/18)	0.80	.36	−0.60	.58
Significant differences within column	1 > 2,3		2 > 1	

Note. *ns* = nonsignificant. Significance value set at $p < .05$. The proportion in the parentheses following each scale name indicates the number of items retained (numerator) from the number of items in the original scale (denominator).

for each scale (for complete item-level information, please consult online Appendixes D through K). IRT analyses estimate two parameters for each item, “alpha” and “beta.” Alpha, which is also referred to as the slope or discrimination parameter, corresponds to the item’s ability to discriminate between individuals and can be analogized to the item’s quality. Beta corresponds to the level of the latent trait that is required for an individual to endorse a given response with a 50% probability. Within intellectual assessment, beta is often analogized as the item’s “difficulty” but within personality and psychopathology assessment it might more accurately be referred to as an item’s “extremity” or “severity.” Also note that the DAPP-BQ/NEO PI-R comparison contains four beta values, while the SNAP/NEO PI-R comparison only has a single beta. This difference is a result of the different response formats employed. The dichotomous items within the SNAP/NEO PI-R comparisons indicate the point along each trait at which the probability of responding “true” begins to exceed that of responding “false.” Conversely, the Likert-type items from the DAPP-BQ/NEO PI-R comparison have five response options and the four beta values correspond to the interval between each of these options. For example, b1 within Table 2 indicates the level of the latent trait at which the likelihood of responding “disagree” becomes higher than that of responding “strongly disagree” to the average item within each scale. Much like the MICs, we calculated the values within Tables 2 and 3 by averaging the beta parameters from each item within each respective scale as well as the standard errors for each beta.

Tables 2 and 3 also include a summary of tests to compare the relative magnitudes of the mean beta parameters. For these tests, we conducted a series of one-way ANOVAs within each domain such that items were treated as cases, each scale’s membership was treated as the independent variable, and beta was the dependent variable. This was followed with Tukey’s post hoc comparisons for each set of scales. Table 2 contains a legend specifying the instances for which the differences in the mean beta parameters were significantly different between two scales ($p < .05$). For example, when comparing the SNAP and NEO PI-R within the domain of emotional instability, the main effect was significant, $F(2, 27) = 4.05$. Post hoc tests revealed that the SNAP self-harm mean (1.41) was significantly higher than the mean value for the NEO PI-R neuroticism items (0.84), but not different from the SNAP dependency mean (1.15).

This finding is readily interpretable for the dichotomous items within the SNAP and NEO PI-R comparisons. The difference in these mean betas indicates that the level of the latent trait required to endorse an item with a 50% probability is significantly higher for the SNAP self-harm items than for the NEO PI-R neuroticism items. Table 3 also indicates that the differences between the mean beta values for NEO PI-R extraversion, the SNAP detachment, and SNAP exhibitionism scales were not significant, $F(2, 23) = .88$. In contrast, the main effect for the SNAP-NEO antagonism ANOVA was significant, $F(4, 46) = 2.60$. Post hoc tests revealed that the SNAP aggression scale had a significantly higher beta than the NEO PI-R agreeableness scale. Finally, the main effect for the SNAP-NEO constraint comparison was significant, $F(2, 34) = 4.38$, and Table 3 indicates that the SNAP Impulsivity scale had a significantly higher beta than did the NEO PI-R conscientiousness scale.

The interpretation of the four beta parameters within the DAPP-NEO comparison in Table 2 is somewhat more complex so we chose to focus on the third beta parameter (i.e., b3), which corresponds to the point at which we dichotomized the NEO items for the SNAP comparison. Within the emotional instability domain, the main effect was significant, $F(6, 133) = 13.38$, and post hoc tests revealed that the DAPP-BQ suicidal ideation scale had a b3 value that was significantly higher than those from all other scales except DAPP-BQ insecure attachment which, in turn, also was significantly higher than the DAPP-BQ anxiety and affective lability scales. The NEO PI-R neuroticism scale was generally lower than the values for most DAPP-BQ scales and these differences were significant for suicidal ideation, submissiveness, identity disturbance, and insecure attachment. Within the introversion domain, the only significant difference found at the third beta was the DAPP-BQ intimacy problems scale having a higher beta than the DAPP-BQ social avoidance scale, $F(3, 27) = 3.44$. The DAPP-NEO comparison was significant, $F(5, 108) = 4.90$ and post hoc tests indicated the agreeableness scale had a mean beta value that was significantly higher than both the DAPP-BQ narcissism and rejection scales. There were no significant differences found within the constraint domain among the third beta values for the DAPP-NEO comparison.

Discussion

A great deal of research has suggested that instruments assessing maladaptive personality traits share a common higher order structure with four of the five FFM domains (Clark, 1993, 2007; Clark et al., 1996; Clark & Livesley, 2002; Livesley, 2003; Markon et al., 2005; O'Connor, 2005; Schroeder et al., 1992; Widiger & Samuel, 2005; Clark et al., in press). However, to date, the research has been confined to exploratory factor analytic studies. In the current study, scales and items from the DAPP-BQ, SNAP, and NEO PI-R were sorted into four higher order domains: emotional instability, antagonism, introversion, and constraint. The items within each domain then were subjected to nonlinear factor analysis to investigate the degree to which the different measures assessed a common, unidimensional latent trait. The results provided evidence that scales assessing normal personality and abnormal personality traits share a common dimensional structure. For example, when the NEO PI-R items assessing extraversion were pooled with those from the SNAP scales of exhibitionism and detachment, the resulting scale appeared to show essential unidimensionality (Stout, 1987), indicating that the items all assess a shared latent construct identified in the current study as introversion. Comparable findings also were obtained when the NEO PI-R scales and those from the DAPP-BQ and SNAP were combined into domains identified herein as emotional instability, antagonism, and constraint. Thus, these findings are consistent with previous evidence that items assessing personality pathology and normal personality traits form at least four unidimensional domains. In addition, the results of the current study go beyond these analyses to apply an IRT perspective to support the hypothesis that the maladaptive traits assessed by the DAPP-BQ and SNAP are extreme versions of general personality structure (Clark, 2007; Livesley, 2005; Widiger & Samuel, 2005).

However, it should be pointed out that, not surprisingly, the DAPP-BQ and SNAP did not operationalize the four latent constructs in precisely the same way. Previous research has indicated that DAPP-BQ and SNAP scales can be understood with respect to four common higher order domains (Clark & Livesley, 2002; Clark et al., 1996; Markon et al., 2005), but the respective scales from each inventory do not define these four domains in precisely the same manner. As a result, content analysis revealed that NEO PI-R items retained with the respective DAPP-BQ and SNAP scales also varied somewhat across the two sets of factor analyses. For example, nearly all the items were retained from the DAPP-BQ/NEO PI-R constraint comparison, making it quite faithful to the original domain as assessed by each instrument. However, the content analysis revealed that the corresponding SNAP/NEO PI-R construct was somewhat narrower and more heavily laden with impulsivity, in part because content assessed in that domain in the SNAP (e.g., the propriety scale) is not well represented in the NEO PI-R. In other cases where the resulting latent trait did differ, this was also a result of the preexisting content within the SNAP or the DAPP-BQ. This finding is not surprising because, despite their similarities, these instruments are not identical in coverage and content. Thus, one would not expect them to define shared latent constructs in precisely the same manner.

However, a strength of the current study is the examination of two different measures of maladaptive personality, rather than relying solely on just one of them. Conducting separate analyses with both measures allows for a contrast that would not have been possible had only one measure been included. By examining the SNAP and DAPP-BQ within in the same manuscript we can consider similarities and differences in their conceptualizations more directly. Nonetheless, future research that administers all of these instruments within the same sample would be useful both to determine whether our findings replicate and also how the domains would be defined when all three measures are considered conjointly.

Scale Comparisons

A considerable body of research has suggested that measures of normal and abnormal personality traits are closely related to one another and involve common, underlying traits (Clark, 2007; Livesley, 2005; Widiger & Samuel, 2005). The IRT analyses allow further explication of this relationship, and indicate that normal and abnormal personality scales occupy different locations on these underlying traits. More specifically, the NEO PI-R generally provided more psychometric information at the lower levels of the latent trait, whereas the DAPP-BQ and SNAP generally provided more information at the higher ends of the underlying trait dimensions. These findings are consistent with the dimensional view of PD, which proposes that PD symptomatology represents not only maladaptive variants of normal personality traits but also extreme (elevated) variants of these same traits (Clark, 2007; Widiger & Samuel, 2005). In this view, for example, the affective lability and self-injurious behaviors associated with borderline PD are more extreme manifestations of the dispositional trait of neuroticism assessed by normal personality inventories (Trull, Widiger, Lynam, & Costa, 2003).

Many prior studies have failed to support a categorical distinction between normal and abnormal personality functioning (e.g., Rothschild, Cleland, Haslam, & Zimmerman, 2003) and have demonstrated meaningful associations between normal and abnormal personality functioning (Samuel & Widiger, 2008; Widiger & Costa, 2002). The current findings provide empirical support for the hypothesis that the primary difference between instruments designed to assess normal versus pathological personality traits is the location along the shared dimensions where they provide the most psychometric information. This indicates that personality pathology instruments provide better fidelity in assessing individuals with high levels of a latent trait, such as antagonism, whereas normal personality instruments provide more discrimination among individuals at the agreeable end of the dimension. The NEO PI-R can contribute to the assessment of personality disorder, such as "borderline" traits (Trull et al., 2003) and psychopathy (Miller & Lynam, 2003), but our IRT analyses suggest that, not surprisingly, measures of personality pathology provide more information regarding maladaptive functioning whereas general personality inventories have greater value within the normal range. In sum, the results presented in Tables 2 and 3, as well as illustrated in Figures 1–3, support the view that personality disorder is a maladaptive expression of normal personality traits.

Nevertheless, a visual inspection of Figures 1–3 suggested that despite significant differences, the curves from these instruments also showed a great deal of overlap. In some cases, such as the DAPP-BQ suicidal ideation scale, the differences were unmistakable and consistent with theoretical expectations. However, in others the differences among curves were less substantial or even nonsignificant (e.g., SNAP exhibitionism). Thus, it appears that despite the differences noted above, there may be overlap as well as distinction between some assessments of normal and maladaptive personality traits. This implies that some scales from the SNAP, DAPP-BQ, and NEO PI-R are providing similar information regarding individuals' standing on the four identified higher order dimensions.

It is important to note that this does not mean, for example, the SNAP exhibitionism scale lacks utility. Rather, because the SNAP was developed expressly to provide lower order, specific trait information (vs. higher order, general factor information) its scales likely are less saturated with the latent trait variance being modeled in these analyses (e.g., introversion/extraversion). To the degree that any item or scale is not assessing the exact latent, higher order construct, its ability to provide psychometric information from an IRT perspective will be limited. Although only those items that loaded sufficiently on the latent trait were retained, it is necessarily true that some items will load more strongly than others. To this extent, there may be items (and thus scales) that will be "favored" or "disfavored" a priori relative to the higher order dimension.

Nonetheless, we do contend that the overlap indicates that the NEO PI-R is not best classified as simply a measure of normal personality functioning. Although it was constructed as a measure of normal personality traits (Costa & McCrae, 1992), it appears that it is better understood as a measure of general personality structure that, in some cases, clearly extends into the realm of abnormal personality functioning. Consider, for instance, the NEO PI-R neuroticism scale. Endorsing items keyed in the direction of depressiveness, anxiousness, self-consciousness, or vulnerability is, in large part, an endorsement of maladaptive personality functioning (e.g., "I am easily frightened," "Sometimes I feel completely worthless," and "At times I have been so ashamed I just wanted to hide"). These items are not appreciably different in context or coverage from respective items from the DAPP-BQ (e.g., "I tend to overreact to minor problems") and the SNAP (e.g., "I haven't made much of my life").

Haigler and Widiger (2001) demonstrated empirically that 98% of the NEO PI-R neuroticism items assess maladaptive personality functioning when keyed in the direction of high neuroticism. Equally important for the purposes of this study, they also found that 83% of the NEO PI-R items keyed in the direction of antagonism (corresponding to DAPP-BQ conduct problems and interpersonal disesteem, see Figure 1) and 90% of the NEO PI-R items keyed in the direction of introversion (corresponding to SNAP detachment, see Figure 2) concern abnormal, maladaptive personality functioning. It is perhaps not surprising then that the IRT curves demonstrated considerable overlap of the NEO PI-R scales with the respective scales from the DAPP-BQ and SNAP, even though the scales were constructed with quite different purposes in mind: The NEO PI-R to assess personality traits evident within the general (normal) population, and the DAPP-BQ and SNAP to assess abnormal personality traits (particularly those underlying the *DSM–IV–TR* personality disorders) within clinical (and more general) populations. In sum, the NEO PI-R may provide more information about the maladaptive range of traits assessed by the DAPP-BQ and SNAP than might be expected from its development and description within the literature as a measure of normal personality functioning. Nevertheless, the IRT analyses do suggest that both the DAPP-BQ and the SNAP are more successful (again, not surprisingly, given the purpose for their development) than the NEO PI-R in covering the highest, most maladaptive range of personality functioning.

Behavioral Specificity

The most distinct findings obtained for either a DAPP-BQ or a SNAP scale were those for DAPP-BQ suicidal ideation. Within Figure 3, the curve for the DAPP-BQ suicidal ideation scale is visibly different from both the NEO PI-R neuroticism scale and even the DAPP-BQ affective lability scale. While most of the curves peak around an average to moderately high level of theta, this MIC peaks well to the right of the figure at a theta of approximately 2.2.

One explanation for this finding is the extremely low endorsement rates for the suicidal ideation items. In fact, the base-rates have at times been so low for this scale that it has been excluded from past factor analyses of the DAPP-BQ (e.g., Larstone et al., 2002). It is perhaps self-evident that suicidal ideation would be an extreme variation of more general

NEO PI-R depressiveness or DAPP-BQ affective lability. It stands to reason that an individual who would endorse an item such as "I have tried to end my life more than once" would be more extreme on the trait of emotional instability than would an individual who would only endorse a NEO PI-R item such as "Sometimes I feel completely worthless." The suicidal ideation items, however, are not only more dysfunctional; they also tend to be more behaviorally specific. For example, two more items from the suicidal ideation scale are "I have taken an overdose when I was very upset" and "I have tried to commit suicide." These items are quite behaviorally specific, relative to simply endorsing the presence of depressed mood or even suicidal ideation. It is perhaps the behavioral specificity of these items that also contributes to the more distinctive locations along the underlying trait dimension. It is also noteworthy that—despite their specificity—they still contain sufficient general trait variance that they were retained in the joint factor analyses. It should also be noted that the parameter estimates for these items are not based upon only a handful of subjects. Even within the community sample there was a notable prevalence of suicidality (e.g., 47 persons indicated that ending their lives seems to be the only way out, and 42 persons indicated that they have tried to kill themselves).

One of the unique strengths of the IRT approach to scale construction and evaluation is its ability to identify items that provide information at a specific location along an underlying dimension (Embretson & Reise, 2000). An ideal assessment of personality structure from the perspective of IRT would be to have items that provide specific assessments at all levels of the trait, analogous to items of an achievement or ability test providing precise discriminations at each point along increasing levels of ability. The findings from the current study clearly indicate that the DAPP-BQ, NEO PI-R, and the SNAP do a good job of assessing the broad range of four primary personality dimensions, although their strengths do lie in somewhat different ranges. Nevertheless, it would be useful for future research to explore the development of items that are specific to each severity level of the latent trait. Minimally, it would be useful to have items whose information curves are specific to the normal range or abnormal range of the trait, respectively, and ideally to have items that provide specific discriminations within each of these ranges.

It is quite possible that such individual items already exist within the scales of the DAPP-BQ, SNAP, or NEO PI-R. However, it was beyond the scope of the current study to report the item response curves for the approximately 1,000 items from these three instruments. Moreover, there is reason to doubt that a large number of such items are present within the existing measures, because previously published IRT analyses of these instruments have not identified an appreciable number of items unique or specific to particular points along the respective trait dimensions (Reise & Henson, 2000; Simms & Clark, 2005). One common thread among these instruments is the reliance on classical test theory (CTT) methodologies for their construction and validation. While this strategy has assuredly produced reliable and valid measures of personality dimensions, it may have also led to the exclusion of the more extreme, and perhaps behaviorally specific, items that may be useful, if not necessary, for specific discriminations along a respective trait dimension. Thus, future research should address this hypothesis by using IRT to evaluate existing items from various personality pathology measures as well as developing new, experimental items to determine whether better discrimination can be obtained along specific trait dimensions. While the development of more behaviorally specific items may increase the measurement range of a scale, it is also possible that these items would evidence weaker loadings with a broad, general factor. For this reason, the potential inclusion of such items must be carefully weighed against the unidimensionality requirement (assuming the goal of measuring the general domain vs. more specific traits). Nonetheless, the potential for behaviorally specific, low base-rate items to increase the measurement range of personality measures appears to be fertile ground for future study.

Limitations

The two samples were each relatively large by traditional standards (i.e., over 600 subjects), but these sample sizes are only adequate for IRT purposes. Additionally, they were community and undergraduate samples that were largely

Caucasian. Therefore, future research that replicates these findings within larger, more ethnically diverse, samples that vary more widely in terms of severity may be useful. Furthermore, the current study employed the self-report methodology exclusively. Although self-report is the most commonly used method within research and clinical practice (Widiger & Boyd, 2009), the accepted gold standard for assessment is the structured interview (Rogers, 2001). Although there is not an a priori reason to believe that other assessment methods would evidence different results (as the SNAP and DAPP-BQ were designed to assess the same information, albeit via self-report), it will nonetheless be important for future research to replicate these findings using structured interviews to assess the *DSM–IV–TR* (or *DSM–V*) personality disorders and the FFM.

In addition, the method we used for testing differences among the betas is limited by its dependence on the number of items within a given domain. In these ANOVA analyses, we treated the items as cases, which makes the detection of significant differences among the beta values much more difficult for domains that retained fewer scales or items. This could be an explanation for why the DAPP-NEO emotional instability comparison (with 140 items) had several significant findings whereas the SNAP-NEO introversion comparison (with 26 items) had none. Perhaps had we included the SNAP "temperament" scales in our analysis, the latter result would have been more similar to the former. Finally, a potential limitation—or at least complication—of this study is the contrasting response formats employed by each instrument. The dichotomous items of the SNAP are consistent with the derivation of IRT technologies within intellectual and academic testing formats that rely almost exclusively on "correct" (vs. incorrect) response options. This property gives them the advantage of being readily interpretable, particularly in the case of the beta parameters. Either a given item (or scale) is more extreme or it is not. However, one benefit inherent to polytomous items, such as those present on the DAPP-BQ and the NEO PI-R, is the further differentiation of individuals with respect to multiple response categories. This same principle may also allow polytomous items to provide greater psychometric information at the extreme levels of the latent traits. The potential influence of response format on IRT analyses is an important area for future research.

Conclusions

This study provided a demonstration, using IRT methods, that scales from personality pathology and general personality instruments can be combined onto a common metric and that the personality pathology items provide more information at extreme levels of these traits than do general personality items. Self-report scales assessing personality pathology and general personality traits were shown to lie along common underlying continua, with the two sets of scales generally differing significantly in terms of their respective locations along the latent trait. Whereas the normal personality scales tended to provide greater information at the lower levels of theta, the personality pathology scales consistently provided more information at the upper levels. This evidence supports both a dimensional conceptualization of personality disorder and the utility of IRT in future instrument development and evaluation.

References

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

Bolt, D. M., Hare, R. D., Vitale, J. E., & Newman, J. P. (2004). A multigroup item response theory analysis of the Psychopathy Checklist-Revised. *Psychological Assessment, 16,* 155–168.

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

Clark, L. A. (2007). Assessment and diagnosis of personality disorder. Perennial issues and an emerging reconceptualization. *Annual Review of Psychology, 58,* 227–257.

Clark, L. A., & Livesley, W. J. (2002). Two approaches to identifying the dimensions of personality disorder: Convergence on the five-factor model. In P. T. Costa, Jr. & T. A. (Eds.) *Personality disorders and the five-factor model of personality* (2nd ed., pp. 161–176). Washington, DC: American Psychological Association.

Clark, L. A., Livesley, W. J., Schroeder, M. L., & Irish, S. L. (1996). Convergence of two systems for assessing specific traits of personality disorder. *Psychological Assessment, 8,* 294–303.

Clark, L. A., Simms, L. J., Wu, K. D., & Casillas, A. (in press). *Manual for the Schedule for Nonadaptive and Adaptive Personality (SNAP-2).* Minneapolis, MN: University of Minnesota Press.

Cooke, D. J., & Michie, C. (1997). An item response theory analysis of the Hare Psychopathy Checklist-Revised. *Psychological Assessment, 9,* 3–14.

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

Costa Jr., P. T., Herbst, J. H., McCrae, R. R., & Siegler, I. C. (2000). Personality at midlife: Stability, intrinsic maturation, and response to life events. *Assessment, 7,* 365–378.

Costa Jr., P. T., & McCrae, R. R. (1985). *The NEO personality inventory manual.* Odessa, FL: Psychological Assessment Resources.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Erlbaum.

Haigler, E. D., & Widiger, T. A. (2001). Experimental manipulation of NEO PI-R items. *Journal of Personality Assessment, 77,* 339–358.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1–55.

Jane, J. S., Oltmanns, T. F., South, S. C., & Turkheimer, E. (2007). Gender bias in diagnostic criteria for the personality disorders: An item response theory analysis. *Journal of Abnormal Psychology, 116,* 166–175.

Jang, K. L., Livesley, W. J., & Vernon, P. A. (2002). The etiology of personality function: The University of British Columbia Twin Project. *Twin Research, 5,* 342–346.

Larstone, R. M., Jang, K. L., Livesley, W. J., Vernon, P. A., & Wolf, H. (2002). The relationship between Eysenck's P-E-N model of personality, the five-factor model, and traits delineating personality dysfunction. *Personality and Individual Differences, 33,* 25–37.

Lilienfeld, S. O., & Andrews, B. P. (1996). Development and preliminary validation of a self-report measure of psychopathic personality traits in noncriminal populations. *Journal of Personality Assessment, 66,* 488–524.

Livesley, W. J. (2001). Conceptual and taxonomic issues. In W. J. Livesley (Ed.), *Handbook of personality disorders: Theory, research, and treatment* (pp. 3–38). New York: Guilford Press.

Livesley, W. J. (2003). Diagnostic dilemmas in classifying personality disorder. In K. A. Phillips, M. B. First, & H. A. Pincus (Eds.), *Advancing DSM: Dilemmas in psychiatric diagnosis* (pp. 153–190). Washington, DC: American Psychiatric Association.

Livesley, W. J. (2005). Behavioral and molecular genetic contributions to a dimensional classification of personality disorder. *Journal of Personality Disorders, 19,* 131–155.

Livesley, W. J., & Jackson, D. (in press). *Manual for the Dimensional Assessment of Personality Pathology–Basic Questionnaire.* Port Huron, MI: Sigma Press.

Markon, K. E., Krueger, R. F., & Watson, D. (2005). Delineating the structure of normal and abnormal personality: An integrative hierarchical approach. *Journal of Personality and Social Psychology, 88,* 139–157.

McCrae, R. R., & Costa, P. T., Jr. (1999). A Five-Factor theory of personality. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (2nd ed., pp. 139–153). New York: Guilford Press.

Miller, J. D., & Lynam, D. R. (2003). Psychopathy and the five-factor model of personality: A replication and extension. *Journal of Personality Assessment, 81,* 168–178.

Mullins-Sweatt, S. N., Jamerson, J. E., Samuel, D. B., Olson, D. R., & Widiger, T. A. (2006). Psychometric properties of an abbreviated instrument of the five-factor model. *Assessment, 13,* 119–137.

O'Connor, B. P. (2005). A search for consensus on the dimensional structure of personality disorders. *Journal of Clinical Psychology, 61,* 323–645.

Reise, S. P., & Henson, J. H. (2000). Computerization and adaptive administration of the NEO PI-R. *Assessment, 7,* 347–364.

Reise, S. P., Smith, L., & Furr, R. M. (2001). Invariance on the NEO PI-R neuroticism scale. *Multivariate Behavioral Research, 36,* 83–110.

Rogers, R. (2001). *Diagnostic and structured interviewing. A handbook for psychologists.* New York: Guilford Press.

Rothschild, L., Cleland, C., Haslam, N., & Zimmerman, M. (2003). A taxometric study of borderline personality disorder. *Journal of Abnormal Psychology, 112,* 657–666.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika, 34*(Suppl. 17), 1–100.

Samuel, D. B., & Widiger, T. A. (2008). A meta-analytic review of the relationships between the five-factor model and *DSM–IV–TR* personality disorders: A facet level analysis. *Clinical Psychology Review, 28,* 1326–1342.

Schroeder, M. L., Wormsworth, J. A., & Livesley, W. J. (1992). Dimensions of personality disorder and their relationships to the Big Five dimensions of personality. *Psychological Assessment, 4,* 47–53.

Simms, L. J., & Clark, L. A. (2005). Validation of a computerized adaptive version of the Schedule for

Nonadaptive and Adaptive Personality (SNAP). *Psychological Assessment, 17,* 28–43.

Smith, L. L., & Reise, S. P. (1998). Gender differences on negative affectivity: An IRT study of differential item functioning on the Multidimensional Personality Questionnaire stress reaction scale. *Journal of Personality and Social Psychology, 75,* 1350–1362.

Stout, W. F. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52,* 589–617.

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika, 55,* 293–325.

Tellegen, A. (1982). *Multidimensional Personality Questionnaire.* Unpublished manuscript, University of Minnesota, Minneapolis.

Thissen, D., Chen, W. H., & Bock, D. (2003). *Multilog 7.03.* Lincolnwood, IL: Scientific Software International.

Trull, T. J., Widiger, T. A., Lynam, D. R., & Costa, P. T., Jr. (2003). Borderline personality disorder from the perspective of general personality functioning. *Journal of Abnormal Psychology, 112,* 193–202.

Waller, N. G. (2002). *WinMFact 2.0.* Minneapolis, MN: Author.

Walton, K. E., Roberts, B. W., Krueger, R. F., Blonigen, D. M., & Hicks, B. M. (2008). Capturing abnormal personality with normal personality inventories: An item response theory approach. *Journal of Personality, 76,* 1623–1648.

Widiger, T.A., & Boyd, S. (2009). Personality disorder assessment instruments. In J. N. Butcher (Ed.), *Oxford handbook of personality assessment* (3rd ed., pp. 336–363). New York: Oxford University Press.

Widiger, T. A., & Costa, P. T., Jr. (2002). Five-factor model personality disorder research. In P. T. Costa, Jr. & Widiger, T. A. (Eds.), *Personality disorders and the five-factor model of personality* (2nd ed., pp. 59–87). Washington, DC: American Psychological Association.

Widiger, T. A., & Samuel, D. B. (2005). Diagnostic categories or dimensions: A question for *DSM-V. Journal of Abnormal Psychology, 114,* 494–504.

Zickar, M. J., Russell, S. S., Smith, C. S., Bohle, P., & Tilley, A. J. (2002). Evaluating two morningness scales with item response theory. *Personality and Individual Differences, 33,* 11–24.