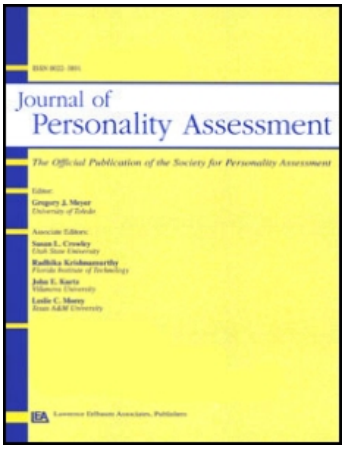
December, 2002

Substance or style? An investigation of the NEO-PI-R validity scales

Leslie C. Morey, *Texas A & M University - College Station*
Brian D. Quigley, *Texas A & M University - College Station*
Charles A. Sanislow
Andrew E. Skodol
Thomas H. McGlashan, et al.



Journal of Personality Assessment
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t775653663

Substance or Style? An Investigation of the NEO-PI-R Validity Scales

Leslie C. Morey; Brian D. Quigley; Charles A. Sanislow; Andrew E. Skodol; Thomas H. McGlashan; M. Tracie Shea; Robert L. Stout; Mary C. Zanarini; John G. Gunderson

Online publication date: 10 June 2010

To cite this Article Morey, Leslie C. , Quigley, Brian D. , Sanislow, Charles A. , Skodol, Andrew E. , McGlashan, Thomas H. , Shea, M. Tracie , Stout, Robert L. , Zanarini, Mary C. and Gunderson, John G.(2002) 'Substance or Style? An Investigation of the NEO-PI-R Validity Scales', Journal of Personality Assessment, 79: 3, 583 — 599
To link to this Article: DOI: 10.1207/S15327752JPA7903_11
URL: http://dx.doi.org/10.1207/S15327752JPA7903_11

JOURNAL OF PERSONALITY ASSESSMENT, *79*(3), 583–599

Substance or Style? An Investigation of the NEO–PI–R Validity Scales

Leslie C. Morey and Brian D. Quigley
Department of Psychology
Texas A&M University

Charles A. Sanislow
Yale University School of Medicine
New Haven, Connecticut

Andrew E. Skodol
Department of Psychiatry
Columbia University and
New York State Psychiatric Institute
New York, New York

Thomas H. McGlashan
Yale University School of Medicine
New Haven, Connecticut

M. Tracie Shea and Robert L. Stout
Department of Psychiatry and Human Behavior
Brown University and
Veterans Affairs Medical Center

Mary C. Zanarini and John G. Gunderson
Department of Psychiatry
Harvard Medical School and
McLean Hospital
Boston, Massachusetts

The Revised NEO Personality Inventory (NEO–PI–R; Costa & McCrae, 1992b) has been criticized for the absence of validity scales designed to detect response distortion. Recently, validity scales were developed from the items of the NEO–PI–R (Schinka, Kinder, & Kremer, 1997) and several studies have used a variety of methods to test their use. However, it is controversial whether these scales are measuring something that is substantive (such as psychopathology or its absence) or stylistic (which might be effortful distortion or less conscious processes such as lack of insight). In this study, we used a multimethod-multitrait approach to examine the validity of these scales in a clinical sample of 668 participants diagnosed with personality disorders or major depression. Using various indicators of both stylistic and substantive variance, confirmatory factor analyses (CFA) suggested that these validity scales measure something that may be conceptually distinct from, yet highly related to, substantive variance in responding.

The Revised NEO Personality Inventory (NEO–PI–R; Costa & McCrae, 1992b) is an inventory designed to assess the five dimensions of personality as described by the Five-factor model (for a review of the Five-factor model, see McCrae & John, 1992; Wiggins, 1996). The five trait dimensions that have emerged from factor analyses of numerous trait terms and various personality inventories have been described as Neuroticism, Extraversion, Openness, Agreeableness, and Conscientiousness. In addition to these five primary personality dimensions, the NEO–PI–R measures six facet subscales that define each dimension and that provide more detailed descriptions of personality characteristics.

Although the Five-factor model has been criticized (Block, 1995; Butcher & Rouse, 1996; Davis & Millon, 1993), the NEO–PI–R has been and continues to be widely used in personality research and in numerous applied contexts. Excellent reliability and stability and sufficient evidence of convergent and discriminant validity of the instrument have been demonstrated by numerous empirical studies (Costa & McCrae, 1992a). Several authors have suggested that the NEO–PI–R may have diagnostic or treatment-related use in clinical settings (Costa & McCrae, 1992a; Piedmont & Ciarrocchi, 1999; Trull, 1992). However, others have questioned the appropriateness of using a measure of "normal" personality to assess psychopathology (Butcher & Rouse, 1996; Clark, 1993b; Coolidge et al., 1994) and the extent to which the NEO–PI–R may be susceptible to response distortion in this population (Ben-Porath & Waller, 1992). A general criticism of the NEO–PI–R is that the instrument's authors failed to include scales intended to detect invalid response sets (Ben-Porath & Waller, 1992). As several empirical studies have demonstrated, the validity of the NEO–PI–R (and the shorter form NEO–Five-Factor Inventory) results can be compromised by deliberate attempts to fake good or fake bad (Bradshaw, 1997; Paulhus, Bruce, & Trapnell, 1995; Topping & O'Gorman, 1997).

Ben-Porath and Waller (1992) suggested that with any self-report measure used in clinical assessment a critical and fundamental step is to determine the validity, or freedom from distortion, of the resultant protocol. They argued that the validity of a psychological test must be evaluated for each individual and that accurate interpretation of the results proceeds from an evaluation of the protocol's degree of validity. Therefore, Ben-Porath and Waller questioned the appropriateness of using the NEO–PI–R in clinical assessment because it does not contain any continuous measures of test validity.

In response to such comments, Costa and McCrae (1992a, 1992c, 1997) have in turn questioned the use of validity scales in personality assessment. They stated that special validity scales were deliberately omitted from the NEO–PI because (a) there is evidence suggesting that, in general, patient self-reports are trustworthy; (b) there is evidence demonstrating that validity scales can be counterproductive and that most social desirability scales are unable to distinguish between individuals who fake good and those who honestly report desirable characteristics; (c) attempts at eliciting cooperation are more likely to improve test validity than are attempts to evaluate, and sometimes make corrections for, protocol invalidity; and (d) as with any assessment device, the NEO–PI–R is not infallible and clinicians should always interpret protocols in the context of supplemental information. Instead, the authors included as a validity check a single item that asks the respondent whether they answered all of the questions honestly and accurately. We would add to their concerns that, particularly in the area of assessing personality pathology, distortion in the negative direction may in fact reflect disordered personality.

Although this issue remains largely unsettled and continues to be ardently debated, researchers have begun to develop methods aimed at detecting response bias and distortion on the NEO–PI–R. For example, Ross, Bailey, and Millis (1997) developed a multivariate function comprised of four facet scales from the NEO–PI–R (impulsiveness, assertiveness, straightforwardness, and dutifulness). Selection of these facets was based on their expected sensitivity and specificity for classifying fake-good and honest protocols. Application of the multivariate model to a sample of college students (Ross et al., 1997) instructed to provide two protocols, one faking good and one responding honestly, correctly classified 86.5% of the fake-good protocols and 88.0% of the honest protocols.

In a study that has stimulated a number of research efforts with regard to the issue of response distortion on the NEO–PI–R, Schinka, Kinder, and Kremer (1997) devised a set of research scales that aim to measure response distortion on the test. These scales assess tendencies to respond randomly (INC; Inconsistency scale) and to present oneself in an overly positive (Positive Presentation Management scale [PPM]) or negative fashion (Negative Presentation Management [NPM] scale). The PPM and NPM validity scales were derived from items on the

NEO–PI–R and selected based on statistical methods and analysis of item content. The INC scale was developed using item pairs that were significantly correlated. Schinka et al. (1997) reported that when participants were instructed to respond in either an honest, positive, or negative manner, significant differences were found in the expected directions on the PPM and NPM scales.

Since the scales were published, a few studies have explored the potential utility of these scales in identifying response distortion. Using a sample comprised of military recruits and college students, Rolland, Parker, and Stumpf (1998) studied group differences on the scales. These authors reasoned that, influenced by the demand characteristics of each administration, a group of military recruits would be more likely to engage in PPM and a group of college students may be more likely to respond randomly. In addition, because NPM and PPM scores are negatively correlated, the college students were also predicted to have greater NPM scores. Findings confirmed these hypotheses, revealing that the military sample had significantly higher scores on PPM than the college student sample, and the college student sample had significantly higher scores on NPM and INC. Furthermore, group membership accounted for 47% of the variance of PPM scores and 16% of the variance of NPM scores.

Support for the use of the Schinka et al. (1997) validity scales was also found in a study by Caldwell-Andrews, Baer, and Berry (2000). After completing the NEO–PI–R under standard instructions, one of three different instructional sets to encourage dissimulation was then given to a sample of college students during a second administration. On the PPM scale, a cutoff score of 22 had an overall hit rate of 79% and a cutoff score of 16 on the NPM scale had an overall hit rate of 85%. Sensitivity, specificity, and positive and negative predictive power were also calculated and revealed respectable probabilities in this sample.

McCrae, Stone, Fagan, and Costa (1998) examined correlations between the Schinka et al. (1997) scales and computed indexes of profile agreement between self-reports and observer reports on the NEO–PI–R for a community sample of married participants. The authors hypothesized that if the scales were identifying systematic distortion on the part of a respondent, the degree of profile agreement with the description provided by an informant would be negatively related to scores on the PPM, NPM, and INC scales. In contrast to these hypotheses, findings did not reveal any consistent relationships for the extent of agreement on any of the five factors or on the total personality profile.

In a similar study, Piedmont, McCrae, Riemann, and Angleitner (2000) argued that studies employing a faking paradigm in which participants are asked to distort their responses are not examining the utility of validity scales in real-world applications and that to assess response bias, external criteria independent of the respondent's self-report must be used. Using a volunteer sample, the authors examined the relationship between self-reports and observer reports of the NEO–PI–R while attempting to control for the suppressor variance of the research

validity scales in the self-reports. They hypothesized that if the validity scales measure response bias, then the correlation between the self-reports and observer reports should be larger when the validity scales are included as suppressor variables. Examination of the zero-order correlations and semipartial correlations between test scores and external criteria demonstrated that in a majority of cases the semipartial correlations were smaller than the zero-order correlations. Piedmont et al. argued that these findings suggest that the validity scales "may actually have substantive content that is related to the criterion" (p. 587).

Although simulation studies have supported the use of the scales, it is still unclear whether the scales could discriminate between honest responding and sources of response distortion in a clinical population. As previously noted, Costa and McCrae (1992a, 1992c, 1997) raised this concern and argued that validity scales are typically unable to distinguish between individuals faking good and individuals honestly reporting desirable characteristics. The converse concern may also be raised in regards to using validity scales with clinical populations—namely, can such validity scales also distinguish between individuals faking bad and individuals honestly reporting undesirable characteristics (i.e., psychopathology)? In both instances, the question is raised as to whether the research validity scales are measuring something substantive (i.e., psychopathology or its absence) or something stylistic (either effortful distortion or something less conscious such as exaggeration or lack of insight). To date, no published studies could be found that examined the utility of these scales in a clinical population.

Consideration of this question suggests that there are three possible interpretations of elevations on the NEO validity scales. First, if these scales are measuring only substantive qualities, they may have very little utility for measuring response distortion in a clinical setting. In this sense, high scores on the NPM would be indicative of psychopathology and psychological distress rather than attempts to falsely report negative characteristics (i.e., malingering). Second, if these scales are measuring only stylistic qualities, they may have demonstrable utility in clinical settings. Therefore, high scores on these indicators of response distortion would provide valid measures of the degree to which an individual may be reporting dishonestly. The third possibility is that these scales may confound both substantive and stylistic qualities and be unable to distinguish between them. In this regard, these measures of response distortion might have some utility but would have poor discriminant validity in clinical settings, suggesting that stylistic tendencies may be inextricably related to substantive qualities, each having an effect on the other. Stylistic tendencies themselves may be representative of personality pathology. An example of this can be found in the clinical literature on depression, which suggests that individuals with psychological depression also have a tendency to exaggerate the negative elements of their experience (e.g., Morey, 1996).

The purpose of this study was to examine the utility of the NPM and PPM scales developed by Schinka et al. (1997) in a sample comprised of clinical participants.

The previously mentioned studies examined the use of these scales with strictly nonclinical populations and the extent to which their findings are generalizable to clinical settings is unknown. This study used a multimethod-multitrait approach (e.g., Campbell & Fiske, 1959) in an effort to explore the construct validity of the NEO scales, examining the convergence of these scores with other indicators of both stylistic and substantive variance. CFA methods were used to explore the adequacy of various models representing the relationship of the Schinka et al. scales to these various indicators to determine whether these scales are best modeled as indicators of response substance or of response style.

METHOD

Participants

Study participants were evaluated as part of a prospective, repeated measures project to examine the longitudinal course of personality disorders (Gunderson et al., 2000). To do this, primarily treatment-seeking participants were sampled for four representative personality disorders (borderline, schizotypal, avoidant, and obsessive–compulsive personality disorders) along with a comparison group meeting criteria for major depressive disorder but with no personality disorder. Three disorders were chosen to represent the three clusters of the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM;* American Psychiatric Association, 1994) and the fourth disorder, obsessive–compulsive, was included because of factor analytic studies suggesting a fourth factor (e.g., Hyler & Lyons, 1988; Kass, Skodol, Charles, Spitzer, & Williams, 1985). Treatment-seeking individuals were targeted so that the results of the study would have real-world application to the individuals who present for treatment (for a detailed description of the study rationale, see Gunderson et al., 2000).

Participants aged 18 to 45 years were recruited primarily from patients seeking treatment at clinical services affiliated with each of the four recruitment sites of the study. The sample was supplemented by participants responding to postings or media advertising for an interview study of personality; such respondents were currently seeking or receiving psychiatric treatment or psychotherapy, or had recently been in psychiatric treatment or psychotherapy. Potential participants were prescreened to determine age eligibility and treatment status or history to assist in excluding patients with active psychosis, acute substance intoxication or withdrawal, a history of schizophrenia-spectrum psychosis (i.e., schizophrenia, schizophreniform, or schizoaffective disorders), or organicity. All eligible participants who began the assessment signed written informed consent after the research procedures had been fully explained. The final cohort for the study was comprised of 668 participants, each assigned to one of five cells: Major depressive disorder

with no personality disorder (n = 97), schizotypal (n = 86), borderline (n = 175), obsessive–compulsive (n = 153), or avoidant (n = 157) personality disorders. The total sample was 64% women, with the largest ethnic groups being White (76%), African American (11%), and Hispanic (9%); for complete demographic information, see Gunderson et al. (2000). The typical rates of Axis II diagnostic co-occurrence were found in the Collaborative Longitudinal Personality Disorders sample (for a detailed description of the diagnostic composition of the sample, see McGlashan et al., 2000).

Assessments

All participants were interviewed by experienced research interviewers with master's or doctoral degrees and extensive training. At baseline, the Structured Clinical Interview for *DSM–IV* (4th ed.; American Psychiatric Association, 1994) Axis I Disorders (SCID–IV; First, Gibbon, Spitzer, & William, 1996) and the Diagnostic Interview for Personality Disorders–IV (Zanarini, Frankenburg, Sickel, & Yong, 1996; median reliability kappas of .69 to .97 for all Axis II disorders; see Zanarini et al., 2000) were among the interview assessments conducted. As part of these assessments, interviewers were asked to complete two ratings scales of particular importance in this investigation. The Global Assessment of Functioning (GAF) Scale, an overall assessment of symptomatic and functional impairment, was rated as described by Axis V of the *DSM–IV*. Also, an item indicating the interviewers' appraisal of the quality of the participant's data (with respect to reliability and accuracy) was also rated; this involved a 5-point Likert-type scale rating with higher scores indicating questionable data quality.

In addition to the interview methods, participants completed self-report instruments including the NEO–PI–R (Costa & McCrae, 1992b) and the Schedule for Nonadaptive and Adaptive Personality (SNAP; Clark, 1993a). The NEO–PI–R (Costa & McCrae, 1992a) was designed to provide a comprehensive assessment of the Five-factor model of personality; these domains include Neuroticism, Extraversion, Openness to Experience, Agreeableness, and Conscientiousness. NEO–PI–R also measures six facet scales that define each of the five domains. The 240 items are answered on a 5-point Likert scale ranging from 1 (*strongly disagree*) to 5 (*strongly agree*). Internal consistency reliabilities for the five domain scales range from .86 to .95; for the facet scales they range from .56 to .81. The temporal stability of the NEO scales have been demonstrated over periods spanning several years and high correlations have been obtained between self-reports and observer ratings (Costa & McCrae, 1992a). In this sample, internal consistency of facets ranged from .58 to .85 (median [*Mdn*] = .75), whereas the domain scales ranged from .87 to .92 (*Mdn* = .89). The PPM and NPM validity scales for the NEO–PI–R were calculated using the proce-

dures outlined by Schinka et al. (1997); internal consistencies of these scales in the study sample was .51 for PPM and .62 for NPM.

The SNAP is a self-report questionnaire designed to assess personality characteristics in both the normal and the abnormal range. There are 12 lower order trait dimensions that load primarily onto one of three higher order factors: Negative Temperament, Positive Temperament, or Disinhibition (each also marked by a corresponding scale). The SNAP has been found to have good internal consistency (*Mdn* α = .81) and good temporal stability over a 2-month period (*Mdn r* = .79; Clark, 1993a). Internal consistency in our study sample was also quite good, with a median of .89 for the higher order temperament scales and median of .84 for the lower order trait scales. Of particular significance in this study, the SNAP also includes a number of validity scales of which three were of particular interest in this study: Desirable Response Inconsistency (DRIN), assessing the tendency to respond based on social desirability rather than the item content; Rare Virtues, which identifies participants who are presenting themselves in a very favorable light; and Deviance, which identifies respondents who are identifying themselves as extremely deviant. The latter two scales yielded respective internal consistency estimates of .56 and .61, respectively, in our sample.

RESULTS

As a first step in the analyses, descriptive statistics for PPM (*M* = 14.90, *SD* = 4.75) and NPM (*M* =13.27, *SD* = 4.75) were calculated in this clinical sample. The mean score for PPM was substantially below (effect size = 1.18 *SD*) community norms provided by Schinka et al. (1997), whereas scores on NPM were substantially above (effect size = 1.18 *SD*) community norms. A total of 23.6% of our sample were 2 *SD*s above Schinka et al.'s community norms for NPM, whereas only 0.2% were 2 *SD*s above norms for PPM. Using impression management group scores reported by Schinka et al. as a guideline, 2.8% of our patient group exceeded the NPM mean score of their negative impression management group, whereas 1.4% of our patients exceeded the PPM mean score of their positive impression management group. These percentages give some estimate of the potential false positive rate of these cutting scores when applied to patient samples.

The zero-order correlations among a series of variables representing potential stylistic and substantive aspects of the reporting of psychiatric symptomatology are presented in Table 1. These variables include four summary indicators of distress and impairment: the GAF score, the SNAP Negative Temperament score, and two aggregated scores reflecting the total number of Axis I diagnoses indicated by the SCID and the total number of Axis II criteria indicated by the Diagnostic Interview for Personality Disorders. For each variable, higher scores suggest greater impairment and distress. Also included are four putative indicators of response styles, the

TABLE 1
Zero–Order Correlations Among Substantive and Stylistic Indicators of Patient Status

Variable	*NPM*	*PPM*	*DRIN*	*Deviance*	*Rare Virtues*	*Negative Temperament*	*GAF*	*Axis I Dx*	*Axis II Sx*	*Reliability*
NPM	1.00									
PPM	–0.40**	1.00								
DRIN	–0.24**	0.15**	1.00							
Deviance	0.54**	–0.38**	–0.29**	1.00						
Rare Virtues	–0.04	0.19**	–0.08	–0.05	1.00					
Negative Temperament	0.24**	–0.44**	–0.12**	0.32**	–0.06	1.00				
GAF	–0.26**	0.20**	0.23**	–0.36**	–0.04	–0.27**	1.00			
Axis I Dx	0.14**	–0.20**	–0.07	0.22**	–0.13**	0.30**	–0.18**	1.00		
Axis II Sx	0.30**	–0.34**	–0.19**	0.42**	0.00	0.48**	–0.40**	0.34**	1.00	
Reliability/ Quality	0.11*	–0.10*	–0.11**	0.21**	–0.01	0.19**	–0.32**	0.09*	0.20**	1.00

Note. N = 641. NPM = Negative Presentation Management scale; PPM = Positive Presentation Management scale; DRIN = Desirable Response Inconsistency scale; GAF = Global Assessment of Functioning Scale; Axis I Dx = total number of Axis I diagnoses; Axis II Sx = total number of Axis II criteria; Reliability/Quality = rated quality of interview information.

*$p < .05$, two tailed. **$p < .01$, two tailed.

quality of information rating provided by the interviewer, and three validity scales from the SNAP: Deviance, Rare Virtues, and DRIN. For each of these scales, extremely high scores would be considered to suggest response distortion. However, because some scales (e.g., Rare Virtues and DRIN) measure distortion in a positive direction and others (e.g., Deviance) measure distortion in a negative direction, some of these variables display inverse relationships. Finally, correlations with the PPM and NPM scales are also included in this table. These correlations reveal that both of the NEO–PI–R validity scales demonstrated significant correlations with nearly every other variable, both stylistic and substantive, measured in the study. The NPM scale demonstrated its largest association with the SNAP Deviance scale (.54), whereas the largest association with PPM was an inverse relationship with the SNAP Negative Temperament scale (–.44).

To test hypotheses more specifically about the nature of the PPM and NPM scales as indicating substantive or stylistic features of responding, CFA methods (Jöreskog, 1969) were utilized to test latent constructs representing each of these hypothetical sources of variance. A confirmatory approach was deemed advantageous over a more purely exploratory approach, as the variables in this study were designed as indicators of either stylistic or substantive sources of variance in assessment. Thus, the analyses were designed to test the hypothesis that this conceptual distinction provides a good fit to observed data. To carry out these analyses, we utilized the Analysis of Moment Structures software (Version 3.6; Arbuckle, 1997). Four alternative models were tested using a full multitrait-multimethod CFA that each specified two assessment method factors—self-report and interview—in addition to specified trait factor(s). The approach was a variant of the preferred "correlated uniqueness" model (Kenny & Kashy, 1992) that allowed a correlation between the method factors as well as the trait factors but specified method trait and error variables as uncorrelated.

Four systematic variants of this model were tested. First, a one-factor solution was tested to examine the fit of a unidimensional model of responding. In this model, all observed variables were considered to reflect the operation of a single latent variable. The fit of this model tested the most parsimonious hypothesis, shedding light on whether the supposedly stylistic and substantive indicators might actually all be measuring the same latent construct, and provided a benchmark comparison for the more elaborate models. Next, two separate two-factor models were tested; one representing the Schinka et al. (1997) scales as grouped with the indicators of response style, and the second representing these scales as grouped with indicators of substantive impairment and distress. Thus, in one model NPM and PPM were considered as indicators of the same latent variable (a putative stylistic variable) as DRIN, Deviance, Rare Virtues, and Reliability/Quality of information; in the alternative model, PPM and NPM were considered as indicators of the same latent variable (a putative substantive variable) as GAF, Negative Temperament, Axis I diagnoses, and Axis II symptoms. Finally, an at-

tempt was made to fit a three-factor solution corresponding to the latent constructs of substantive impairment and a division of the response style indicators into two nested components, positive impression management and negative impression management. All models shared identical observed variables, and most of the alternative models were nested, thereby facilitating comparisons using statistical significance tests. For each of these comparison models, we specified independence of error terms. For those models involving more than one latent variable, we allowed the latent variables to be correlated.

Goodness-of-fit indexes for the one-factor model suggested a reasonable fit to the data, $\chi^2(24, N = 668) = 72.26$ (nonnormative fit index [NFI] = .93; comparative fit index [CFI] = .951; root mean square error of approximation [RMSEA] = .061). Next, the alternative two-factor models, alternatively considering the PPM and NPM scales as stylistic or substantive sources of variance, were tested. The fit indexes for the PPM/NPM as stylistic model suggested an appreciable improvement on the single-factor model, $\chi^2(23, N = 668) = 46.37$ (NFI = .955; CFI = .976; RMSEA = .043); this improvement was a significant increment over the one-factor model, $\chi^2(1, N = XX) = 25.89$, $p < .01$. In contrast, the PPM/NPM as substantive model represented no improvement on the single-factor model, $\chi^2(23, N = 668) = 71.51$ (NFI = .93; CFI = .951; RMSEA = .062), which was not a significant improvement over the one-factor model, $\chi^2(1, N = 668) = 0.75$, $p > .05$. Finally, the three-factor solution, in which PPM and NPM were considered as stylistic indicators and these indicators were divided into positive and negative forms of distortion, provided minimal improvement, $\chi^2(21, N = 668) = 42.17$ (NFI = .960; CFI = .978; RMSEA = .040) on the analogous two-factor model that did not separate stylistic components into positive and negative features, $\chi^2(2, N = 668) = 4.21$, $p > .05$.

Based on the increments in quality of fit, the PPM/NPM as stylistic solution appeared to offer the best fit of our hypothesized models. This model and the resulting standardized loading estimates are presented in Figure 1. By all fit indexes, the overall model provided a good fit to the data. Significant path coefficients in this model, as determined by a critical ratio test ($p < .05$, two-tailed), are italicized in the figure.

DISCUSSION

This study examined the pattern of relationships between the Schinka et al. (1997) NEO–PI–R validity scales and various other indicators of stylistic and substantive variance. These relationships indicated that scores on these scales display sizable associations with measures of global functioning as well as with putative measures of response validity. The association with validity and functional measures tended to hold across both self-report and interviewer-based methods of assessment.

A variety of different covariance structural models of these relationships were examined for fit. A model hypothesizing correlated stylistic and substantive fac-

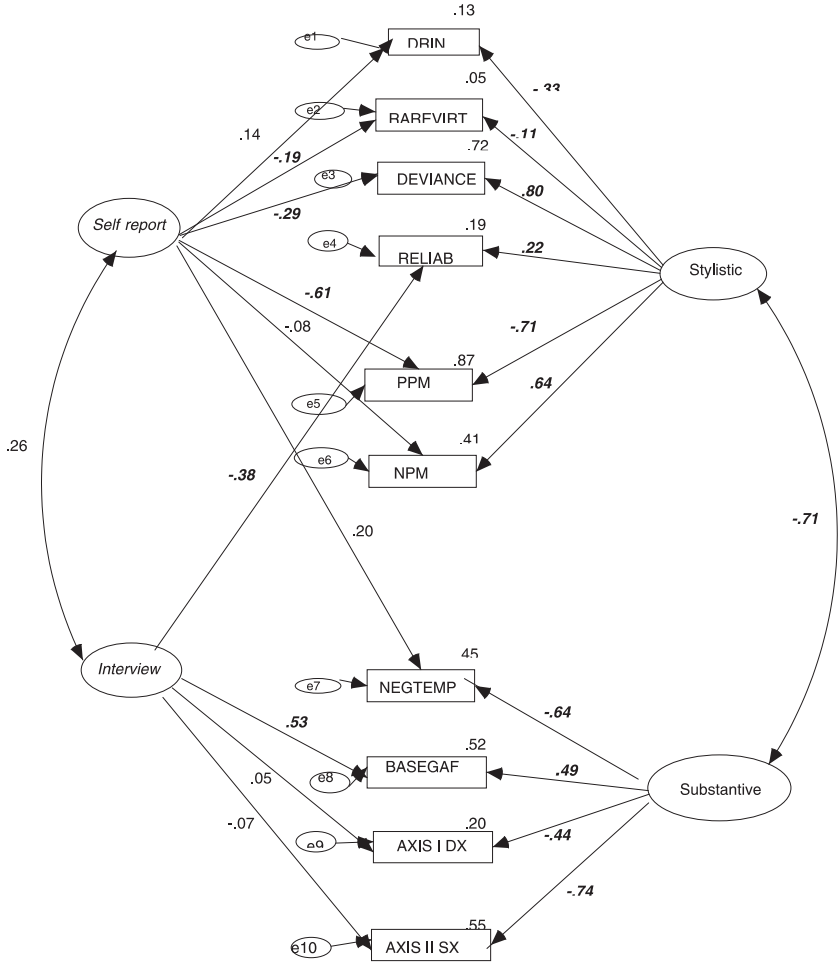


FIGURE 1 Structural model of stylistic and substantive indicators.

tors representing the NEO validity scales as part of the stylistic factor, with additional response method factors, provided a relatively good fit to the data. This model, shown in Figure 1, has a number of potentially important implications for the use of the NEO validity scales and perhaps for the use of such scales in general.

One important finding was the relatively poor fit of a single-factor modeling of these data. This finding suggests that it is better to conceptualize stylistic and substantive sources of variance as conceptually distinct, even in a sample in which there is little apparent motivation for a distorted self-presentation. For example,

when a respondent reports marked dysfunction, it is probably best to consider that there may be both stylistic and substantive contributions to this report and that it may be profitable to think of these two sources of variance separately. In the arena of assessing personality pathology, this seems particularly apt. Thus, indicators of distortion, such as NPM or PPM, may indeed have their place in the armament of personality assessors. Interestingly, in this study of clinical participants, measures of positive and negative distortion did not appear to be distinct but rather as opposing poles of the same construct. This finding could in part be due to the nature of the sample, which provided reports in a context (a naturalistic research study) with very little incentive for any form of distortion. In contexts in which motivations for distortion might be more powerful or more complex (such as preemployment screening or dissimulation studies), the positive and negative forms could prove to be more differentiable.

However, another very important result from Figure 1 is that the correlation between the latent stylistic and substantive factors was estimated at a sizable .71. In other words, nearly half of the variance of these hypothetical constructs is shared. As a result, although the factors may be conceptually distinct, they are by no means independent, at least as found in clinical samples. This is consistent with the controversies surrounding the use of validity scales in general and has a number of theoretical and practical implications for the use of such scales in clinical settings. First, the type of response distortion measured by indicators such as PPM and NPM is heavily intertwined with the respondent's substantive functional status. Viewing oneself and one's world in a negative fashion may be an integral part of many mental disorders; a positive and perhaps even repressive self-view may be an integral part of mental health. In light of the magnitude of this estimated relationship between substantive and stylistic factors, it would be a serious error in interpretation to consider elevations on scales such as NPM and PPM as de facto evidence of effortful response distortion (even without considering that response distortion in the negative direction itself may qualify as personality pathology). Simply because experimental simulations can produce elevations on these scales does not imply that the conditions these studies attempt to simulate are the sole, or even primary, reasons that these scales may be elevated. Nonetheless, it should be noted that construction of similar measurement models in a sample of individuals in which motivation to distort might be much stronger (as in forensic, custody, or preemployment evaluations) might lead to much greater separation of stylistic and substantive components, suggesting a profitable direction for future investigation in this area.

Despite this marked relationship, it appears advantageous to consider the stylistic and substantive constructs as distinct conceptually. In other words, there is more to mental health than a tendency to deny minor flaws, and there is more to mental disorder than a tendency to report deviant behaviors. The stylistic factors of self-presentation may still prove to be valuable in assisting in the interpretation of

patient self-reports of strengths and weaknesses, as these reports could be distorted in positive or negative ways. However, it is important to recognize that this distortion is by no means independent of mental health and functional status and indeed may be an integral part of some of the clinical constructs explored in this study. For example, a tendency to distort experience in a negative way may be an important element of borderline personality (e.g., Kurtz & Morey, 1998), whereas defensively minimizing subjective distress may be an integral part of obsessive–compulsive personality. Future research might profitably focus on that part of the stylistic element of responding that appears to be independent of the substantive elements, as a better understanding of that component of stylistic variance could improve the efficiency of validity scales.

The model we presented here may also help clarify why psychometric efforts to correct for stylistic aspects of response variance tend to meet with limited success. Historically, efforts such as the Minnesota Multiphasic Personality Inventory's (MMPI; Hathaway & McKinley, 1943) K correction (Archer, Fontaine, & McCrae,1998) or forced-choice alternatives that attempt to equate items for social desirability (e.g., Edwards, 1957) have not fared well as correction strategies in clinical settings. The results obtained here support the observation by Piedmont et al. (2000) that attempts to correct NEO–PI–R profiles through the use of scales like PPM or NPM are likely to decrease rather than increase validity, as the magnitude of the correlation between the stylistic and substantive factors suggests that any correction is quite likely to remove valid variance from the NEO–PI–R domain scores.

A number of limitations to this study should be considered as limits to generalizability and as important directions for additional research. First, although multimethod in nature, this study was necessarily limited in the number and nature of measures of response style that were obtained. For example, a number of indicators of response distortion have been recently explored that seem to be considerably less associated with criterion variance, such as the MMPI–2 (Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989) F(p) scale or the PAI's (Morey, 1991) Rogers discriminant function (Rogers, Sewell, Morey, & Ustad, 1997). Further exploration of these indicators using modeling strategies similar to those in this study would be interesting to determine if (a) such indicators appear to reflect a coherent latent factor and (b) whether any such factor is indeed less associated with substantive variance. In addition, additional research examining the influence of such factors as diagnosis or situational context is recommended; our study was limited to four specific personality disorders tested in a routine research context. Different patterns might emerge in other diagnoses (e.g., antisocial or narcissistic personality) or other contexts (e.g., forensic or employment screening) that have different types of expectations for impression management.

The use of validity scales for the NEO–PI have sparked some controversy in the literature, with some advocating the potential utility of the scales to identify dis-

torted responding and others dismissing such efforts as heavily confounded with true personality status. The results of this study shed some light on this controversy that illustrates the complexities involved in personality assessment, suggesting that both viewpoints contain important accurate elements. These results suggest that stylistic and substantive factors in self-presentation can successfully be modeled as distinct factors but that these factors are highly related. Rather than interpreting response style scales as indicators of effortful deception, or conversely as simply indicators of functional status, these scales reflect a presentational style that itself may be an integral part of mental health. Future research could be profitably directed at gaining a greater understanding of both halves of the variance in these stylistic scales: that which is fundamentally related to mental health and that which is not (and their relationship to one another). Important research along these lines (e.g., Paulhus & Reid, 1991) has been directed at the positive forms of distortion; the results obtained here underscore the need for similar work directed at the more negative stylistic forms.

ACKNOWLEDGMENTS

This article was supported by NIMH Grants R10 MH 50837, 50838, 50839, 50840, and 50850. This article was reviewed and approved by the Publications Committee of the Collaborative Longitudinal Personality Disorders Study.

REFERENCES

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Arbuckle, J. L. (1997). AMOS Users Guide (Version 3.6) [Computer software]. Chicago: SPSS.

Archer, R. P., Fontaine, J., & McCrae, R. R. (1998). Effects of two MMPI–2 validity scales on basic scale relations to external criteria. *Journal of Personality Assessment, 70,* 97–102.

Ben-Porath, Y. S., & Waller, N. G. (1992). "Normal" personality inventories in clinical assessment: General requirements and the potential for using the NEO Personality Inventory. *Psychological Assessment, 4,* 14–19.

Block, J. (1995). A contrarian view of the five-factor approach to personality description. *Psychological Bulletin, 117,* 187–215.

Bradshaw, S. D. (1997). Impression management and the NEO Five-Factor Inventory: Cause for concern? *Psychological Reports, 80,* 832–834.

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *MMPI–2: Minnesota Multiphasic Personality Inventory–2: Manual for administration and scoring*. Minneapolis: University of Minnesota Press.

Butcher, J. N., & Rouse, S. V. (1996). Personality: Individual differences and clinical assessment. *Annual Review of Psychology, 47,* 87–111.

Caldwell-Andrews, A., Baer, R. A., & Berry, D. T. R. (2000). Effects of response sets on NEO–PI–R scores and their relations to external criteria. *Journal of Personality Assessment, 74,* 472–488.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81–105.

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

Clark, L. A. (1993b). Personality disorder diagnosis: Limitations of the five-factor model. *Psychological Inquiry, 4,* 100–104.

Coolidge, F. L., Becker, L. A., DiRito, D. C., Durham, R. L., Kinlaw, M. M., & Philbrick, P. B. (1994). On the relationship of the five-factor model to personality disorders: Four reservations. *Psychological Reports, 75,* 11–21.

Costa, P. T., & McCrae, R. R. (1992a). Normal personality in clinical practice: The NEO Personality Inventory. *Psychological Assessment, 4,* 5–13.

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

Costa, P. T., & McCrae, R. R. (1992c). Reply to Ben-Porath and Waller. *Psychological Assessment, 4,* 20–22.

Costa, P. T., & McCrae, R. R. (1997). Stability and change in personality assessment: The Revised NEO Personality Inventory in the year 2000. *Journal of Personality Assessment, 68,* 86–94.

Davis, R. D., & Millon, T. (1993). The five-factor model for personality disorders: Apt or misguided? *Psychological Inquiry, 4,* 104–109.

Edwards, A. L. (1957). *The social desirability variable in personality assessment and research.* New York: Dryden.

First, M. B., Gibbon, M., Spitzer, R. L., & William, J. B. W. (1996). *Structured Clinical Interview for* DSM–IV *Axis I Disorders (SCID–I).* CITY: Biometrics Research Department, New York State Psychiatric Institute.

Gunderson, J. G., Shea, M. T., Skodol, A. E., McGlashan, T. H., Morey, L. C., Stout, R. L., et al. (2000). The Collaborative Longitudinal Personality Disorders study: Development, aims, design, and sample characteristics. *Journal of Personality Disorders, 14,* 300–315.

Hathaway, S. R., & McKinley, J. C. (1943). *The Minnesota Multiphasic Personality Inventory*. Minneapolis: University of Minnesota Press.

Hyler, S. E., & Lyons, M. (1988). Factor analysis of the *DSM–III* personality disorder clusters: A replication. *Comprehensive Psychiatry, 29,* 304–308.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 34,* 183–202.

Kass, F., Skodol, A. E., Charles, E., Spitzer, R. L., & Williams, J. B. W. (1985). Scaled ratings of *DSM–III* personality disorders. *American Journal of Psychiatry, 142,* 627–630.

Kenny, D. A., & Kashy, D. A. (1992). Analysis of the multitrait-multimethod matrix by confirmatory factor analysis. *Psychological Bulletin, 112,* 165–172.

Kurtz, J. E., & Morey, L. C. (1998). Negativism in evaluative judgments of words among depressed outpatients with borderline personality disorder. *Journal of Personality Disorders, 12,* 351–361.

McCrae, R. R., & John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of Personality, 60,* 175–215.

McCrae, R. R., Stone, S. V., Fagan, P. F., & Costa, P. T. (1998). Identifying causes of disagreement between self-reports and spouse ratings of personality. *Journal of Personality, 66,* 285–313.

McGlashan, T. H., Grilo, C. M., Skodol, A. E., Gunderson, J. G., Shea, M. T., Morey, L. C., et al. (2000). The Collaborative Longitudinal Personality Disorders study: Baseline Axis I/II and II/II diagnostic co-occurrence. *Acta Psychiatrica Scandinavica, 102,* 256–264.

Morey, L. C. (1996*). An interpretive guide to the Personality Assessment Inventory.* Odessa, FL: Psychological Assessment Resources.

Paulhus, D., & Reid, D. B. (1991). Enhancement and denial in socially desirable responding. *Journal of Personality & Social Psychology. 60,* 307–317.

Paulhus, D. L., Bruce, M. N., & Trapnell, P. D. (1995). Effects of self-presentation strategies on personality profiles and their structure. *Personality and Social Psychology Bulletin, 21,* 100–108.

Piedmont, R. L., & Ciarrocchi, J. W. (1999). The utility of the Revised NEO Personality Inventory in an outpatient, drug rehabilitation context. *Psychology of Addictive Behaviors, 13,* 213–226.

Piedmont, R. L., McCrae, R. R., Riemann, R., & Anglietner, A. (2000). On the invalidity of validity scales: Evidence from self-reports and observer ratings in volunteer samples. *Journal of Personality and Social Psychology, 78,* 582–593.

Rogers, R., Sewell, K. W., Morey, L. C., & Ustad, K. L. (1997). Detection of feigned mental disorders on the Personality Assessment Inventory: A discriminant analysis. *Journal of Personality Assessment, 67,* 629–640.

Rolland, J. P., Parker, W. D., & Stumpf, H. (1998). A psychometric examination of the French translations of the NEO–PI–R and NEO–FFI. *Journal of Personality Assessment, 7,* 269–291.

Ross, S. R., Bailey, S. E., & Millis, S. R. (1997). Positive self-presentation effects and the detection of defensiveness on the NEO–PI–R. *Assessment, 4,* 395–408.

Schinka, J. A., Kinder, B., & Kremer, T. (1997). Research validity scales for the NEO–PI–R: Development and initial validation. *Journal of Personality Assessment, 68,* 127–138.

Topping, G. D., & O'Gorman, J. G. (1997). Effects of faking set on validity of the NEO–FFI. *Personality and Individual Differences, 23,* 117–124.

Trull, T. J. (1992). *DSM–III–R* personality disorders and the five-factor model of personality: An empirical comparison. *Journal of Abnormal Psychology, 101,* 553–560.

Wiggins, J. S. (1973). *Personality and prediction: Principles of personality assessment.* Reading, MA: Addison-Wesley.

Wiggins, J. S. (Ed.). (1996). *The five-factor model of personality: Theoretical perspectives.* New York: Guilford.

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.

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

Leslie C. Morey
Department of Psychology
Texas A&M University
230 Psychology Building
4235 TAMU
College Station, TX 77843–4235
E-mail: lcm@psyc.tamu.edu

Received November 27, 2001
Revised April 30, 2002