Personality and Social Psychology Review
14(3) 332–346

DOI: 10.1177/1088868310361240
http://pspr.sagepub.com


How Should the Internal Structure of Personality Inventories Be Evaluated?

Christopher J. Hopwood[1] and M. Brent Donnellan[1]

Abstract

Personality trait inventories often perform poorly when their structure is evaluated with confirmatory factor analysis (CFA). The authors demonstrate poor CFA fit for several widely used personality measures with documented evidence of criterion-related validity but also show that some measures perform well from an exploratory factor analytic perspective. In light of these results, the authors suggest that the failure of these measures to fit CFA models is because of the inherent complexity of personality, issues related to its measurement, and issues related to the application and interpretation of CFA models. This leads to three recommendations for researchers interested in the structure and assessment of personality traits: (a) utilize and report on a range of factor analytic methods, (b) avoid global evaluations regarding the internal validity of multiscale personality measures based on model fit according to conventional CFA cutoffs, and (c) consider the substantive and practical implications of model modifications designed to improve fit.

Keywords

personality assessment, factor analysis, construct validity

> Every scientist in the back of his mind takes it for granted that even the best theory is likely to be an approximation to the true state of affairs.
>
> Paul E. Meehl (1990, p. 113)

> Since all models are wrong the scientist must be alert to what is importantly wrong. It is inappropriate to be concerned about mice when there are tigers abroad.
>
> George E. P. Box (1976, p. 792)

The evaluation of the psychometric properties of psychological inventories is a critical element in the process of psychological science because these properties underpin the validity of the scientific study of thoughts, feelings, and behaviors. Many of the psychometric developments that have contributed to the evolution of psychological measurement have occurred in the context of research on personality traits. Broadly speaking, the psychometric viability of personality trait inventories is assessed in terms of their construct validity. In this context, Loevinger (1957) went so far as to equate the construct validity of personality tests with the plausibility of personality theories, so that tests of the validity of such measures also represent tests of their underlying theories. She noted that demonstrating construct validity is a multifaceted enterprise that includes showing (a) that the content of measurement instruments corresponds to theoretical content, (b) that the internal structure of instruments corresponds to the conceptual structure of existing theory, and (c) that the empirical network of relations between scores on instruments and criterion variables is consistent with theory. This article is concerned with the criteria and methods that contemporary personality researchers use to evaluate the second domain of construct validity described by Loevinger: internal structure (also see Steger, 2006).

Internal Structure and the Common Factor Model. The building blocks of personality inventories are the actual items that typically assess fairly specific and narrow thoughts, feelings, or behaviors. Items (e.g., "I often feel blue") are grouped together with other items of similar content into specific scales that are thought to assess a single dimension of personality (e.g., "depression"). In many cases, related scales are thought to cluster into broader personality trait dimensions (e.g., "neuroticism"). Internal structure refers to the orderliness of the actual clustering of related elements, within and across theoretically substantive dimensions.

Internal structure is important for both practical and theoretical reasons. Practically, internal structure increases confidence in the usefulness of summary scores. If an inventory is designed to measure a single construct but it actually measures several (i.e., the actual items do not orderly cluster

[1]Michigan State University, East Lansing

Corresponding Author:
Christopher J. Hopwood, Michigan State University, Department of Psychology, East Lansing, MI 48824-1116
Email: hopwood2@msu.edu

into a single dimension), it can be difficult to interpret the meaning of both the total score based on those items and any predictor–criterion relations involving that summary score. The structure of a personality measure also has theoretical implications, as highlighted by Loevinger. For example, debates continue as to whether optimism and pessimism are separate but correlated dimensions or whether they form a single continuum (e.g., Maydeu-Olivares & Coffman, 2006) and whether global self-esteem meaningfully splits into self-competence and self-liking dimensions (e.g., Tafarodi & Milne, 2002). Likewise, the recovery of three, four, five, six, or seven basic dimensions from omnibus personality inventories has implications for discussions regarding optimal representations of higher order personality traits.

Internal structure is commonly evaluated by some form of factor analysis. Factor analytic procedures represent a class of statistical tools that have evolved over the past 100 years (Cudeck & MacCallum, 2007) with multiple applications in personality and other forms of psychological assessment (Briggs & Cheek, 1986; Goldberg & Velicer, 2006; Thompson, 2004). Researchers often make distinctions between exploratory factor analysis (EFA) and confirmatory factor analysis (CFA); however, both are instantiations of the common factor model (Thurstone, 1947). The underlying principle of the common factor model is that shared variability among manifest variables (e.g., scales on a personality inventory) can be attributed to the presence of a smaller set of common but unobserved variables. Researchers interested in personality structure are therefore keenly interested in determining the number and nature of these unobserved or latent factors.

Psychometric experts frequently argue that EFA is "primarily a data-driven approach, whereas CFA is theoretically grounded" (Byrne, 2005, p. 17). EFA approaches are considered "exploratory" because no explicit a priori assumptions need to be made regarding the number of common (or latent) factors that give rise to associations between measured indicators (i.e., items or scales) or how strongly those indicators should load on the unobserved factors. CFA approaches, on the other hand, are considered "confirmatory" because the researcher must specify on an a priori basis the number of common factors and identify which indicators have meaningful loadings on the stipulated latent factors. Although an even stricter approach to the confirmation of a hypothesized internal structure whereby the researcher specifies the precise loadings of indicators on latent factors is possible, this is rarely done in practice. Instead, researchers specify which loadings should be fixed to zero (to indicate no association between an indicator and a latent variable) and which loadings should be estimated from the available data. Furthermore, several practices in CFA, such as model modifications without theoretical justification or the common but typically unexplored existence of equally well fitting alternative models, suggest that strong and unqualified descriptors such as "confirmatory" are potentially inappropriate (Breckler, 1990).

Nevertheless, EFA generally provides less stringent tests of model viability than does CFA. For instance, the adequacy of an EFA result is typically judged by fairly subjective criteria such as the interpretability of the factor solution in the light of preexisting knowledge of the constructs in question and the usefulness of a particular solution. Notably, a more rigorous criterion for the adequacy of an EFA solution involves testing whether a given solution cross-validates in a new sample (see Guadagnoli & Velicer, 1991). In contrast, the adequacy of a CFA model specification typically involves more systematic judgment as it can be formally evaluated based on the discrepancy between the theoretically implied pattern of covariation between the measured variables and their actual observed pattern of covariation. A number of "goodness-of-fit" measures (e.g., the comparative fit index [CFI], Tucker–Lewis index [TLI], and root mean square error of approximation [RMSEA]) as well as the very stringent χ^2 test of exact fit are typically used to judge model adequacy based on this underlying discrepancy.[1]

The actual practice of conducting factor analytic studies involves a number of important decisions and includes several steps requiring the analyst's judgment (e.g., Byrne, 2001; Fabrigar, Wegener, MacCallum, & Strahan, 1999; Floyd & Widaman, 1995; Goldberg & Velicer, 2006; Russell, 2002; Steger, 2006; Thompson, 2004). As a consequence, both EFA and CFA require a reasonable amount of skill, training, and experience to implement competently, and the literature is replete with examples of questionable uses of both techniques (for reviews, see Fabrigar et al., 1999; Russell, 2002). Nonetheless, factor analysis holds an important place for most researchers concerned with evaluating the structural validity of personality trait measures, and researchers often regard CFA as the "gold standard" technique when it comes to evaluating internal structure (e.g., Anderson & Gerbing, 1988; Thompson, 2004; but also see Brannick, 1995; Goldberg & Velicer, 2006; Lee & Ashton, 2007). This preference likely stems from the apparent correspondence between CFA techniques and a hypothesis-driven, deductive approach to science. In CFA, researchers specify the anticipated factor structure of a measure based on preexisting theory and then evaluate that structure using real data, just like researchers specify falsifiable hypotheses and then conduct systematic investigations to test those hypotheses in the context of general scientific inquiry.

Concerns About CFA in Personality Assessment. Despite the intuitive appeal of CFA techniques, concerns have been raised about how CFA studies are conducted and interpreted in practice. Most notably, the adequacy of model fit is open to energetic debate among researchers who use criteria other than the χ^2 test of exact fit (e.g., see *Personality and Individual Differences*, Vol. 42, Special Issue 5 for a discussion of this issue). As Bentler and Bonett (1980) observed, the

χ^2 test will tend to reject models that are "trivially misspecified" if the sample size is large. This places researchers in an unenviable position because larger samples are generally regarded as better than smaller samples. Accordingly, researchers typically interpret a subset of the myriad indices of close fit to supplement and often supplant the evaluation of the χ^2 test. The downside of this strategy is that it creates the potential for researchers to marshal evidence selectively and conclude that an instrument has a valid or invalid structure depending on which rules of thumb are used to evaluate specific goodness-of-fit indices (Marsh, Hau, & Wen, 2004). As such, researchers, reviewers, and editors might base publication decisions, in part, on how they attend to different markers of model fit.

In fact, several authors (e.g., Marsh et al., 2009; McCrae, Zonderman, Costa, Bond, & Paunonen, 1996) have noted that it is relatively easy to show that a simplified CFA measurement structure usually fails to fit personality data. The issue of poor model fit can be clearly demonstrated by considering putatively unidimensional personality measures. A number of instruments are designed to assess a single, relatively narrow, construct such as life satisfaction, global self-esteem, or depression. CFA models can be used to test the hypothesis of unidimensionality in such measures by evaluating the fit of a measurement model whereby all items load on a single common factor. Implicit in this test of unidimensionality is the concept of local independence, or the idea that manifest indicators are unrelated to each other when controlling for the common factor (Hattie, 1985). The adequacy of a single-factor model for single attribute measures is frequently rejected in practice. For example, Slocum-Gori, Zumbo, Michalos, and Diener (2009) tested whether a single factor fit the five items of the Satisfaction With Life Scale (SWLS; Diener, Emmons, Larsen, & Griffin, 1985) and found that the χ^2 was 32.93 ($df = 5$, $N = 410$), whereas the RMSEA was .117, values that are thought to indicate poor fit by most conventions. Quilty, Oakman, and Risko (2006) tested a single-factor model for the 10 items on the Rosenberg Self-Esteem Scale (RSES; Rosenberg, 1965) and also found poor fit ($\chi^2 = 487.01$, $df = 35$, $N = 503$; RMSEA = .160).

This situation also extends to the evaluation of the structure of omnibus personality inventories using CFA techniques. For example, Church and Burke (1994) and McCrae et al. (1996) both failed to confirm the structure of the widely used NEO Personality Inventory (NEO-PI) or Revised NEO Personality Inventory (NEO-PI-R) in separate investigations. Similar exact fit failures have been reported for these and other Big Five inventories (e.g., Borkeneau & Ostendorf, 1990; Donnellan, Oswald, Baird, & Lucas, 2006; Gignac, Bates, & Jang, 2007; Vassend & Skrondal, 1995). These kinds of results, which are particularly notable in light of the influence of factor analytic methods on the development of the Big Five model of personality, have led some researchers to express seemingly pessimistic perspectives on the theoretical status and utility of the Big Five and the NEO-PI (e.g., Vassend & Skrondal, 1997).

The "Henny Penny" Problem. The observation that many personality measures typically fail exact fit CFA criteria can create what we call a "Henny Penny" problem. Henny Penny problems occur when researchers interpret negative CFA results as implying the need to call into question the meaningfulness of all previous studies using particular measures, including those demonstrating other kinds of validity (e.g., content, criterion-related). These sometimes exaggerated claims are like the concerns of Henny Penny, who lamented that the sky was falling in the proverbial story. Such potential exaggerations can create serious and often unnecessary doubts about the integrity of large literatures. Consider again the construct of life satisfaction. It is common for applied researchers to sum or average the items that constitute the SWLS for use in subsequent statistical analyses. A researcher might exploit the finding that a unidimensional factor model displayed rather poor fit to actual data in a particular sample to call into question the vast literature on life satisfaction that is based on this instrument.

A major reason why we are skeptical of the "sky is falling" conclusion from many CFA studies is that some of the failures of CFA approaches are often understandable on methodological grounds. At the most basic level, it is very difficult to write "perfect" items for assessing personality. This means that items unavoidably tap additional if substantially minor sources of variation (e.g., Slocum-Gori et al., 2009). Such minor factors often create a smattering of correlated residuals in an item-level CFA and generate overall model misfit when not explicitly included in the analysis (see, e.g., Marsh et al., 2009). For instance, two items on the RSES explicitly make reference to social comparisons ("I am able to do things as well as most other people" and "I feel that I'm a person of worth, at least on an equal plane with others"), whereas the other eight items do not explicitly invoke such comparisons. It would not be surprising to find that those two items exhibited residual covariation above and beyond their association because of a general self-esteem factor.

Moreover, personality measures are sensitive to other methodological artifacts stemming from item wording, such as when all of the negatively worded items on an inventory share variance above and beyond a general factor, in violation of the local independence assumption (e.g., DiStefano & Motl, 2009; Marsh, 1996; Quilty et al., 2006).[2] In light of these concerns, researchers have started to become accustomed to notions of "essential" unidimensionality, or the idea that a set of items assesses one dominant latent attribute despite the presence of minor "secondary" factors (Slocum-Gori et al., 2009). Indeed, Hattie (1985) cautioned that it "may be unrealistic to search for . . . sets of [purely] unidimensional items" (p. 159). Given this situation, model misfit might represent an unpleasant consequence of the

complicated nature of personality and its assessment (Goldberg & Velicer, 2006).

Related explanations for model misfit exist when CFA techniques are applied to personality scales in the hopes of testing the higher order structure of personality trait inventories such as the NEO-PI-R (Costa & McCrae, 1992). First, many lower order scales have cross-loadings on multiple factors, although the cross-loadings are typically "minor" from the perspective of EFA studies (e.g., less than .30). These cross-loadings may not reflect measurement problems per se but rather the tendency for many practically important aspects of personality to be located interstitially between broad factors (Ashton, Lee, Goldberg, & de Vries, 2009; Goldberg, 1993; Hofstee, de Raad, & Goldberg, 1992). For example, several facets of the NEO-PI-R (Costa & McCrae, 1992) are linked with aspects of self-control although these facets are aligned with different broader domains (e.g., impulsiveness is associated with Neuroticism, self-discipline is associated with Conscientiousness, and excitement seeking is associated with Extraversion). Such complicated relations are not accommodated by the very restrictive assumptions about structure that are commonly specified in CFA models. Indeed, the most commonly specified kinds of CFA models are called independent cluster models (see Marsh et al., 2009) because they specify that each indicator is associated with only one common factor. However, Goldberg and Velicer (2006) noted that there are "few factor univocal items [as] most items [have] secondary factor loadings of substantial size" (p. 230). The consequence of failing to include cross-loadings in a CFA model is that they are then assumed to be zero and any true deviation from zero contributes to model misfit (see Ashton & Lee, 2007). Unfortunately, specifying all of the relevant minor loadings on an a priori basis appears to be extremely difficult.

Second, correlated residuals between lower order scales within a broad dimension (e.g., depression and vulnerability facet scales within the Neuroticism domain on the NEO-PI-R) are also likely to occur if two scales are more similar to one another than they are to the other facets of the same dimension. In other words, two facets might share some additional correlation above and beyond their shared associations because of their respective links with a common factor. The existence of these residual correlations will contribute to model misfit if they are fixed to zero. Unfortunately, researchers often have limited systematic insight into this source of misfit as these are not traditionally part of EFA approaches to scale development and refinement, approaches that were historically used as the starting point for creating many omnibus inventories.

All in all, several issues may account for the difficulties that occur in attempting to specify exact fitting CFA models for omnibus personality trait inventories that go beyond the empirical adequacy of the measures themselves. Collectively, these kinds of considerations have led some researchers to question the relevance of CFA studies for personality psychology. McCrae et al. (1996, p. 553) made the analogy to aeronautical engineering and claimed that few would call for the grounding of airplanes if an engineer conducted simulations suggesting that such machines cannot fly. Rather, they claimed, most people would suggest that the simulations were flawed. More recently, Lee and Ashton (2007) suggested that personality "researchers should be cautious of [CFA]" (p. 437). Despite these cautionary notes, CFA studies are quite pervasive and influential in the personality assessment literature.

Empirical Illustrations. As it stands, tension continues to exist between the general sentiment that CFA approaches are the "gold standard" approach for evaluating the internal structure of measures and concerns about the limited utility of very restrictive CFA techniques for evaluating personality instruments. In light of this tension, we believe that is it worthwhile to reexamine issues of model fit and model adequacy with respect to EFA and CFA approaches for evaluating the internal structure of personality trait inventories. We believe that such reanalyses will promote an open and frank discussion that could lead to more nuanced perspectives regarding the evaluation of the internal structure of personality measures. With this goal in mind, we evaluated the internal structure of seven personality trait inventories using both CFA and EFA methods. We selected measures (a) that are widely used by researchers and practitioners, (b) whose content is understood to map to a higher order structure, and (c) which have shown acceptable and similar levels of criterion-related validity (Grucza & Goldberg, 2007). The overarching purpose of these analyses was to demonstrate how both EFA and CFA methods assess the internal structure of these well-known and well-regarded measures. This information should provide an empirical starting point for continued discussions regarding the appropriate role of factor analytic approaches for evaluating personality trait inventories.

Method

We chose seven multiscale instruments that were administered to the Eugene Springfield Community Sample (ESCS; see Grucza & Goldberg, 2007, for sample details) and that have a hierarchical structure consisting of lower order scales that are thought to cohere into higher order dimensions.[3]

1. The fifth edition of Cattell's 16PF (Conn & Rieke, 1994; $N = 680$) comprises 185 items with three response options and yields 16 primary factor scale scores that can be combined to represent five higher order factors.
2. The Six-Factor Personality Questionnaire (6fpq; Jackson, Paunonen, & Tremblay, 2000; $N = 714$) has 108 items with a 5-point response scale and yields 18 scale scores that serve as the lower order components of six higher order factors.

3. The California Psychological Inventory (CPI; Gough & Bradley, 1996; N = 792) includes 462 true–false items. We scored its 20 Folk scales and 11 Special Purpose scales (not including 2 response style scales) to correspond to analyses by Gough and Bradley that identified five higher order factors.
4. The HEXACO Personality Inventory (Lee & Ashton, 2004; N = 734) has 192 items with a 5-point response scale that are scored as 24 facets that can be organized as six higher order domains.
5. The Hogan Personality Inventory (HPI; Hogan & Hogan, 1995; N = 742) includes 206 true–false items with 41 lower order homogeneous item clusters and seven higher order scales.
6. The Multidimensional Personality Questionnaire (MPQ; Tellegen & Waller, 2008; N = 733) has 276 true–false items with 11 content scales. Of these scales, 10 are thought to cohere into either a three- (MPQ 3) or four-factor (MPQ 4) model (see Tellegen & Waller, 2008). The remaining content scale, Absorption, appears to measure a relatively distinct dimension of personality. This dimension was not included in confirmatory analyses, whereas it was included for some exploratory analyses, as described below.
7. The NEO-PI-R (Costa & McCrae, 1992; N = 857) comprises 240 items with a 5-point response scale and yields 30 facets associated with five higher level domain scores.

A previous article (Grucza & Goldberg, 2007) showed that each of these measures had similarly impressive validities in predicting an array of theoretically relevant outcome criteria. For example, the average cross-validated multiple Rs for these instruments for predicting behavioral acts (e.g., friendliness, creativity) was .50 (SD = .02, range = .47–.52) at the level of higher order factors or traits and .52 (SD = .02, range = .50–.55) for lower order constructs. These values for observer-reported descriptions were .46 (SD = .05, range = .37–.52) and .46 (SD = .05, range = .41–.52) at the higher and lower order levels, respectively. For clinical indicators, these values were .41 (SD = .08, range = .27–.53) and .42 (SD = .07, range = .29–.52). Accordingly, these inventories provide an interesting test case in that they all have demonstrated similar utility in predicting outcomes, and thus the present illustration has the potential to illuminate the limited correspondence between internal and external validity, as suggested by Loevinger (1957).

Analyses

We assessed the internal validity of each of these measures by modeling the structure of higher and lower order scales using both CFA and EFA techniques. CFAs were modeled with maximum likelihood (ML) estimation in AMOS 17.0. We also replicated a subset of these analyses using Mplus 5.2 and found no substantial differences in the results. Latent variable covariances were all freely estimated (i.e., models were oblique), and measurement paths were specified based on prior analyses or theoretical descriptions in the original validation materials. In cases where previous research suggested that instruments did not have simple structure (e.g., the CPI Social Presence scale had meaningful loadings on the Ascendance, Communality, and Originality factors; see Gough & Bradley, 1996, pp. 60-64), we allowed measured variables to load onto multiple factors. Given controversies regarding which goodness-of-fit statistics should be used and which thresholds should demarcate acceptable fit (Hu & Bentler, 1998; Marsh et al., 2004), we used the following "liberal" criteria: TLI greater than .90, CFI greater than .90, and RMSEA less than .10. These cutoffs are more or less consistent with the methodological "urban legend" surrounding the evaluation of model fit in covariance structure modeling contexts (e.g., Lance, Butts, & Michels, 2006).

Some authors have recommended alternatives to such global indices for evaluating CFA fit, such as using information about expected parameter changes and modification indices to identify and evaluate the consequences of model misspecifications (Saris, Satorra, & van der Veld, 2009). Thus, we also explored the ability of modification indices to enhance CFA fit given previous work suggesting that achieving acceptable fit while remaining faithful to the canonical structure of instruments described in the test manuals might be difficult (e.g., Church & Burke, 1994). However, in light of concerns regarding the use of modification indices to improve fit (MacCallum, 1986), we regarded these analyses as a demonstration rather than as an attempt to identify "the" structure of a given personality inventory. Some have even argued that such post hoc attempts are actually better served using EFA techniques rather than CFA approaches (see Browne, 2001). Given these considerations, model modification was pursued with only one instrument, the NEO-PI-R, which was selected based on its historical role in controversies involving CFA techniques applied to omnibus inventories (e.g., McCrae et al., 1996; Vassend & Skrondal, 1997). We computed modification indices for successive models and freed implicated paths or covariances until no indices were greater than 10 and then conducted a second series of modifications until no indices were greater than 5. We were primarily concerned with three results from these analyses. First, we were interested in how many modifications (and resulting losses in degrees of freedom) would be necessary to achieve acceptable fit. Second, we were interested in the degree to which an optimal model would cross-validate. To cross-validate modified models, we conducted the modification search in a random half of the ESCS sample and then tested the fit of the modified models in the other half of the sample. Finally, we were interested in the substantive implications of these modifications. In particular, we evaluated

the degree to which identified paths or covariances made theoretical sense.

In terms of the EFA analyses, we first used identical factoring and rotation methods as in the original validation data (typically the test manuals) and extracted factors based on their theoretical structure for EFA analyses (in some cases this meant that we conducted a principal components analysis rather than a common factor EFA). We extracted the number of factors anticipated by theory for each instrument. We then conducted additional EFA analyses using ML estimation as implemented in Mplus 5.2. Conducting an EFA using ML estimation procedures provides a number of goodness-of-fit indices that are useful for evaluating the quality of a given factor solution, especially in terms of deciding on the appropriate number of factors (see Brown, 2006, p. 29; Hoyle & Duvall, 2004). The fit of these EFA models estimated with ML also provides a benchmark level of misfit for a model that specified a number of factors because all potential cross-loadings are included in the analysis such that the remaining sources of misfit arise from correlated residuals and/or specification errors related to the number of latent factors (see Mulaik, 2010, pp. 475-476).

We further examined pattern coefficients in a more descriptive manner. We reasoned that most factor analysts desire structures in which scales have strong convergent associations with a single primary factor but generally weak divergent associations with nonprimary factors. Thus, we evaluated how well each of the inventories met these ideals using varying standards for convergent and divergent associations. Specifically, we assessed the number of pattern coefficients for each instrument that were convergent (those with theoretically anticipated associations) and divergent (those that were not theoretically anticipated to associate) at three magnitudes (.20, .30. and .40). For these analyses, we again used the EFA methods from the original studies.

Last, we conducted several sets of analyses designed to examine the generalizability of the internal structure of the personality measures. First, we split the ESCS data randomly for each instrument and computed Tucker congruence coefficients (Guadagnoli & Velicer, 1991; Tucker, 1951; Wrigley & Neuhaus, 1955) and factor pattern Pearson correlations (Louks, Hayne, & Smith, 1989; Teel & Verran, 1991) across the two subsamples for each factor from each instrument. Congruence coefficients are often considered "good" when they exceed .95 and "fair" when they are between .85 and .94 (Lorenzo-Seva & ten Berge, 2006; cf. Chan, Ho, Leung, Chan, & Yung, 1999), and thus these values served as our interpretative benchmarks. For these analyses, we used unweighted least squares extraction and promax oblique rotations and computed congruence coefficients based on pattern weights to address any concerns over limitations imposed by using potentially suboptimal approaches to factoring the inventories (e.g., the use of principal components analysis; Widaman, 2007).

Table 1. Confirmatory Factor Analysis Fit Indicators for Seven Omnibus Personality Assessment Measures

Instrument	χ^2	(*df*)	TLI	CFI	RMSEA	90% CI
16PF	739.58	(74)	.66	.76	.12	(.11–.13)
6fpq	814.08	(120)	.70	.79	.09	(.08–.10)
CPI	Inadmissible					
HEXACO	2353.92	(237)	.59	.65	.11	(.11–.11)
HPI	Inadmissible					
MPQ 3	385.56	(31)	.52	.67	.13	(.11–.14)
MPQ 4	282.44	(27)	.60	.76	.11	(.10–.13)
NEO-PI-R	5296.01	(395)	.57	.61	.12	(.12–.12)

Note: TLI = Tucker–Lewis index; CFI = comparative fit index; RMSEA = root mean square error of approximation; 6fpq = Six-Factor Personality Questionnaire; CPI = California Psychological Inventory; HPI = Hogan Personality Inventory; MPQ = Multidimensional Personality Questionnaire; NEO-PI-R = Revised NEO Personality Inventory. All χ^2 tests were significant ($p < .001$). For the 16PF, we compared Eugene Springfield Community Sample data to results from a principal components analysis on correlation matrix from the manual. Because of its cross-loadings with several higher order factors, absorption was not included in MPQ models. Fit was decremented modestly when absorption was included. Four- and three-factor models have been described for the MPQ, so both were modeled.

We also conducted a variant of Procrustes rotation (Barrett, 1986; Hoelzle & Meyer, 2009; McCrae et al., 1996) using the target rotation specification in Mplus 5.2. Here the goal was to see how well the EFA solution for a particular sample approximated a "target" solution that was designated in advance. Put differently, we supplied the targets for a given instrument and then assessed how well the final solution approximated that target. The degree of approximation was assessed with congruence coefficients (see McCrae et al., 1996). We used two different sets of target matrices reflecting varying levels of understanding about the intended structure of the inventories. The first set of target matrices reflected binary codes depicting theoretical simple structure (i.e., 0s and 1s). The second set of target matrices consisted of coefficients from other large samples, typically derived from test manuals.[4]

Results

CFAs. Table 1 shows the fit of the models using CFA approaches for each inventory.[5] None approached acceptable fit by the conventions typically applied to the evaluation of covariance models. In fact, two models were inadmissible: a negative error variance was estimated for the CPI and HPI analyses produced a nonpositive definite latent variable covariance matrix (similar errors were encountered using Mplus). As discussed above, we also conducted a series of model modifications on a random half of the sample in an effort to improve the fit of the NEO-PI-R.[6] The original model had 395 degrees of freedom; 61 modifications were necessary to achieve a model with no modification indices greater than 10, resulting in 334 degrees of freedom. This model had the

Table 2. Goodness of Fit Statistics for Exploratory Factor Analysis Models

Instrument	χ^2	(*df*)	TLI	CFI	RMSEA	90% CI
16PF	264.33	(50)	.82	.93	.08	(.07–.09)
6fpq	154.06	(60)	.93	.97	.05	(.04–.06)
CPI	3057.67	(320)	.85	.90	.10	(.10–.11)
HEXACO	560.85	(147)	.87	.93	.06	(.06–.07)
HPI	1680.86	(554)	.81	.87	.05	(.05–.06)
MPQ 3	180.69	(18)	.62	.85	.11	(.10–.13)
MPQ 4	75.04	(11)	.75	.94	.09	(.07–.11)
NEO-PI-R	1637.07	(295)	.84	.89	.07	(.07–.08)

Note: TLI = Tucker–Lewis index; CFI = comparative fit index; RMSEA = root mean square error of approximation; 6fpq = Six-Factor Personality Questionnaire; CPI = California Psychological Inventory; HPI = Hogan Personality Inventory; MPQ = Multidimensional Personality Questionnaire; NEO-PI-R = Revised NEO Personality Inventory.

following goodness-of-fit statistics: $\chi^2_{(334)} = 640.205$ ($p < .001$); CFI = .953, TLI = .939, RMSEA = .046. However, the model did not show acceptable fit on cross-validation: $\chi^2_{(334)} = 1103.543$ ($p < .001$); CFI = .873, TLI = .834, RMSEA = .073. To achieve a model with no modification indices greater than 5, it was necessary to conduct 36 further modifications, and this model achieved good fit: $\chi^2_{(298)} = 323.804$ ($p > .10$); CFI = .996, TLI = .994, RMSEA = .014. However, the model again exhibited relatively poor to marginal fit on cross-validation: $\chi^2_{(298)} = 903.118$ ($p < .001$); CFI = .900, TLI = .854, RMSEA = .069.

These analyses show that substantial alterations to the basic structure of the NEO PI-R were necessary even with a purely empirical effort to improve model fit. Furthermore, despite these alterations and the presumed similarity of the cross-validation sample in terms of demographic and other characteristics, the modified model did not cross-validate even according to liberal conventions. We were also interested in the nature of the parameters that were freed to improve fit. Many of these parameters made conceptual sense. For instance, the largest initial modification index suggested the need to free the regression path from the Agreeableness facet of compliance to the Neuroticism facet angry hostility, suggesting an association between these aspects of personality. It is intuitive that angry people also tend to be less compliant. However, many other modification indices suggested associations for facets that were not obviously related to one another (e.g., the Agreeableness modesty facet with the Openness fantasy facet).

EFAs. We then turned to an evaluation of the instruments from an EFA perspective using the strategies outlined above. As a transition from CFA analyses, we first evaluated model fit for the exploratory models as estimated in Mplus. Here we extracted the number of common factors associated with the higher order structure of the model (Table 2). Some of these indices were within an acceptable range for several measures, although only one, the 6fpq, had acceptable fit across all indicators other than the highly sensitive χ^2. No other measure achieved a TLI greater than .90. Four of the measures (16PF, 6fpq, HEXACO, and MPQ 4) had CFI values greater than .90, and all but two (CPI and MPQ 3) had RMSEA values less than .10.[7] These findings suggest that these personality inventories tend to have what would conventionally be regarded as mediocre fit even with the relatively unrestricted EFA model. It is useful to recall that correlated residuals contribute to model misfit in this context, and thus Table 2 suggests that these are a considerable source of difficulty for finding well-fitting models for personality trait inventories.

Table 3. Percentage of Convergent and Divergent Pattern Coefficients Consistent With Theoretical Models at Varying Levels of Specificity

Instrument	Criterion	% Convergent	% Divergent
16PF	0.2	95	57
	0.3	86	86
	0.4	86	100
6fpq	0.2	100	95
	0.3	100	98
	0.4	100	100
CPI	0.2	100	44
	0.3	100	75
	0.4	88	86
HEXACO	0.2	100	87
	0.3	100	98
	0.4	100	100
HPI	0.2	77	83
	0.3	71	90
	0.4	18	95
MPQ 3	0.2	100	90
	0.3	98	98
	0.4	92	98
MPQ 4	0.2	100	82
	0.3	98	95
	0.4	95	98
NEO-PI-R	0.2	100	81
	0.3	100	91
	0.4	99	98

Note: 6fpq = Six-Factor Personality Questionnaire; CPI = California Psychological Inventory; HPI = Hogan Personality Inventory; MPQ = Multidimensional Personality Questionnaire; NEO-PI-R = Revised NEO Personality Inventory. Analyses were based on unweighted least squares promax pattern coefficients. For the convergent column, coefficients > criterion/total expected coefficients; for the divergent column, coefficients < criterion/total expected coefficients. 16PF Self-Control factor did not reflect the anticipated weights in Eugene Springfield Community Sample (ESCS) data, so coefficients were based on first four factors. The HPI Ambition factor did not reflect anticipated weights in ESCS data, so coefficients were computed based on coefficients from the other six factors. The MPQ absorption was not factored because of ambiguities with regard to its factor loadings. Table 13.3 from Tellegen and Waller (2008) was used for the hypothesized pattern coefficients for three- and four-factor models.

Table 3 shows the number of convergent and divergent pattern coefficients that were of the expected magnitude at

Table 4. Congruence Coefficients for Personality Factors in Randomly Halved Samples

	Factor					
Instrument	1	2	3	4	5	6
16PF	Extraversion	Anxiety	Tough-Mindedness	Independence	Self-Control	
Tucker	.99	.86	.66	.63	.64	
Pearson	.99	.89	.63	.74	.70	
6fpq	Extraversion	Agreeableness	Methodicalness	Independence	Openness	Industriousness
Tucker	.98	.96	.95	.99	.97	.94
Pearson	.98	.98	.95	.98	.97	.93
HEXACO	Honesty Humility	Emotionality	Extraversion	Agreeableness	Conscientiousness	Openness
Tucker	.99	.97	.98	.99	.97	.98
Pearson	.99	.97	.97	.98	.97	.98
MPQ 4	Negative Emotionality	Agency	Communion	Constraint		
Tucker	.96	.96	.97	.98		
Pearson	.96	.94	.98	.96		
MPQ 3	Negative Emotionality	Positive Emotionality	Constraint			
Tucker	.91	.75	.91			
Pearson	.96	.65	.85			
NEO-PI-R	Neuroticism	Extraversion	Openness	Agreeableness	Conscientiousness	
Tucker	.99	.98	.98	.98	.99	
Pearson	.98	.97	.98	.98	.99	

Note: 6fpq = Six-Factor Personality Questionnaire; MPQ = Multidimensional Personality Questionnaire; NEO-PI-R = Revised NEO Personality Inventory. Absorption was not factored for the MPQ.

three cutoffs: .20, .30, and .40. More specifically, two values were computed: the percentage of pattern coefficients that were expected, by theory, to be meaningful that were above these cutoffs (% convergent) and the percentage of pattern coefficients that were expected not to be meaningful that were below these cutoffs (% divergent). By this relatively liberal standard, several instruments performed reasonably well. For instance, 100% of the convergent pattern coefficients were greater than .40 for the 6fpq, and 100% of the divergent pattern coefficients were less than .40; in fact, 95% of divergent pattern coefficients were even less than .20. Notably, for most instruments (with the exception of the HPI), strong convergent pattern coefficients appeared easier to achieve than low cross-loadings. For instance, two instruments had 100% convergence across all three values (6fpq, HEXACO), and the CPI, MPQ, and NEO-PI-R also performed reasonably well. However, several divergent percentages were relatively low, and only two models (6fpq and MPQ 3) had percentages greater than 90% for divergent coefficients at all three levels. Given the differential rates across convergent and divergent coefficients, the criterion of .40 tended to yield the best overall hit rate. For four instruments (6fpq, HEXACO, MPQ 3, and NEO-PI-R), both rates were greater than 90% at this level.[8]

Table 4 shows the congruence coefficients for personality factors from each instrument computed within random halves of ESCS data. We were unable to recover the structure reported in the CPI and HPI manuals and therefore did not compute congruence coefficients for these instruments. Among the other instruments, only the NEO-PI-R and HEXACO achieved congruence coefficients that exceeded the cut score of .95 for strong similarity across all factors. By a more relaxed criterion of .85, the MPQ 4 and 6fpq showed acceptable factor structure generalizability, whereas coefficients for the MPQ 3 and 16PF were somewhat lower.

Table 5 shows congruence coefficients for instruments rotated to binary codes indicating pattern coefficients for scales that are theoretically relevant (1) and nonrelevant (0) for each factor. We were unable to make binary decisions regarding scale factor pattern coefficients for the CPI or 16PF based on the level of detail provided in their manuals, so congruence coefficients for these measures were not considered. None of the measures showed congruence coefficients greater than .94 across all factors, and the HEXACO was the only instrument with all coefficients greater than .84. Table 6 shows congruence coefficients for instruments rotated to factor pattern coefficients from previous samples. This method specifies consistent or known cross-loadings (e.g., NEO-PI-R impulsiveness on Neuroticism and Conscientiousness factors) and thus reflects a more sophisticated understanding of inventory structures than the binary approach. Accordingly, these coefficients tended to be higher than those from the binary analyses: All coefficients were greater than .95 for the NEO-PI-R and HEXACO, and all were greater than .84 for the MPQ 3.

Table 5. Congruence Coefficients for Personality Factors Following Procrustes Rotations to Binary Matrices

	Factor						
Instrument	1	2	3	4	5	6	
6fpq	Extraversion	Agreeableness	Methodicalness	Independence	Openness	Industriousness	
Tucker	.91	.94	.93	.90	.90	.86	
Pearson	.90	.93	.91	.88	.88	.83	
HEXACO	Honesty Humility	Emotionality	Extraversion	Agreeableness	Conscientiousness	Openness	
Tucker	.97	.92	.92	.97	.94	.96	
Pearson	.93	.88	.90	.87	.91	.92	
HPI	Adjustment	Ambition	Sociability	Interpersonal Sensitivity	Prudence	Inquisitiveness	Learning Approach
Tucker	.76	.71	.59	.71	.69	.74	.74
Pearson	.69	.65	.54	.66	.64	.69	.69
MPQ 4	Negative Emotionality	Agency	Communion	Constraint			
Tucker	.82	.81	.84	.91			
Pearson	.78	.78	.83	.91			
MPQ 3	Negative Emotionality	Positive Emotionality	Constraint				
Tucker	.93	.84	.93				
Pearson	.91	.83	.93				
NEO-PI-R	Neuroticism	Extraversion	Openness	Agreeableness	Conscientiousness		
Tucker	.86	.82	.92	.82	.87		
Pearson	.88	.77	.90	.79	.86		

Note: 6fpq = Six-Factor Personality Questionnaire; HPI = Hogan Personality Inventory; MPQ = Multidimensional Personality Questionnaire; NEO-PI-R = Revised NEO Personality Inventory. Absorption was not factored for the MPQ.

Discussion

Despite widespread use and considerable evidence for the criterion-related validity of the personality measures we examined (Grucza & Goldberg, 2007), they all tended to disappoint using the common conventions for adjudicating CFA model fit. Moreover, even a purely empirical modification search failed to yield a model for the NEO-PI-R that would effectively cross-validate by common conventions for evaluating CFA model fit. By contrast, several instruments performed reasonably well by more relaxed criteria associated with EFA techniques (e.g., congruence coefficients), consistent with our "eyeball" tests of the factor pattern matrices. Overall, these results raise important questions about how the internal structure of personality inventories should be evaluated.

Implications. One potential reading of these findings is that all of the personality measures we examined are seriously deficient in terms of fidelity to their underlying theories. Taken to its extreme, this position would imply that researchers invested in personality trait constructs should return to the drawing boards and refine existing measures or develop new measures that pass existing criteria for structural validity. Such a conclusion would be quite dramatic given that these instruments represent some of the most commonly used tools in personality psychology. Furthermore, this conclusion would have the potential to bring individual differences research to a virtual standstill. As each of the measures investigated here has demonstrated acceptable and even rather strong criterion-related validity (Grucza & Goldberg, 2007), arguing that they are not useful from a purely practical standpoint makes little sense. We are therefore reluctant to endorse this reading of the results.

An equally extreme conclusion in the opposite direction would suggest that strict evidence for replicable internal structure using factor analytic techniques is not essential for construct validity. Lykken (1971) foreshadowed this opinion by arguing that

> the logic of the factor analytic model is inappropriate for the structure of naturally occurring organic systems, the principle of simple structure is in conflict with known facts in biological science, and factor analytically derived personality variables have not been shown to possess reality and usefulness outside of the factor analytic context. (p. 161; also see Brannick, 1995; Cloninger, 2008; Gough & Bradley, 2002)

Table 6. Congruence Coefficients for Personality Factors Following Procrustes Rotations to Target Matrices

	Factor					
Instrument	1	2	3	4	5	6
16PF	Extraversion	Anxiety	Tough-Mindedness	Independence	Self-Control	—
Tucker	.94	.95	.90	.97	.98	.82
Pearson	.94	.94	.90	.97	.98	.83
HEXACO	Honesty Humility	Emotionality	Extraversion	Agreeableness	Conscientiousness	Openness
Tucker	.97	.99	.98	.99	.98	.98
Pearson	.97	.98	.97	.99	.98	.98
MPQ 3	Negative Emotionality	Positive Emotionality	Constraint			
Tucker	.92	.98	.96			
Pearson	.91	.97	.97			
NEO-PI-R	Neuroticism	Extraversion	Openness	Agreeableness	Conscientiousness	
Tucker	.98	.97	.98	.99	.99	
Pearson	.98	.97	.98	.99	.99	

Note: MPQ = Multidimensional Personality Questionnaire; NEO-PI-R = Revised NEO Personality Inventory. Pattern matrices were from the following sources: 16PF—Table 1.3 from Conn and Rieke (1994); HEXACO—raw data from a student sample provided by M. Ashton; MPQ—Table 13.3 from Tellegen and Waller (2008); NEO-PI-R from McCrae, Zonderman, Costa, Bond, and Paunonen (1996), appendix. For the 16PF, a sixth factor emerged in Conn and Rieke with a strong loading on the Reasoning scale but was not interpreted at the higher order level; that factor was also considered here (denoted as —). For the MPQ, Absorption was factored because it was factored by Tellegen and Waller (2008). For this instrument, a four-factor model did not converge when rotated to the target matrix.

However, this position is problematic to the extent that it could lead to an "anything goes" mentality in personality assessment. Indeed, we contend that personality assessment has improved over time, in part because of the increasingly sophisticated and thoughtful use of factor analytic methods (see, e.g., Briggs & Cheek, 1986; Jackson, 1971).

Given what is currently known about personality theory and personality trait assessment, we advocate a middle-ground reading of these results in lieu of either of these extreme positions. On one hand, it is important to acknowledge that the measures studied here have demonstrated their utility in a number of applications and across hundreds of studies. In light of our concerns about the "Henny Penny" problem, we do not think that personality trait research should be dismissed based on the uniform failure of omnibus personality inventories to fit very restrictive CFA models. On the other hand, we acknowledge the possibility for improvement in terms of how well multiscale personality trait measures conform to their hypothesized structure, and we believe that the present results could motivate research leading to measures with increasingly well-defined structures.

To be clear, we strongly endorse the perspective that a reliable and theoretically coherent structure is an important psychometric property of multiscale instruments. To the extent that tests amount to operationalized theories (Loevinger, 1957) and demonstrations of the coherence and relations among concepts are valuable for explicating the nature of personality, understanding the structure of measurement tools is crucially important for personality science. We therefore strongly object to an "anything goes" mentality in personality assessment. Furthermore, we believe that both CFA and EFA methods are useful for evaluating a number of test properties and should continue to be used to improve existing measures and inventories. Our overall perspective is that these results should draw attention to the need for researchers to think more critically than has sometimes been the case in the past about how common factor methods are used to evaluate the structure of personality inventories in practice.

Considerations for the Critical Evaluation of Personality Trait Structure. In particular, we believe that several considerations are relevant when evaluating the structure of personality inventories. First, researchers need to think more carefully about the kinds of statistical models that are being specified when researchers evaluate the structure of omnibus inventories with CFA approaches. The current understanding of personality theories and inventories may not yet be sufficiently refined to initially specify the kinds of models that tend to perform well in CFA contexts (i.e., independent cluster models; see Goldberg & Velicer, 2006; Marsh et al., 2009). Our experiences in conducting these analyses suggest to us that it will be quite difficult to specify an "exact" CFA model for a given existing personality inventory even in view of a good deal of knowledge about its general structure. Most notable are correlated residuals that seem to contribute to a considerable amount of misfit (see Table 2; Gignac et al., 2007; Marsh et al., 2009); however, researchers often have little insight into these parameters. Moreover, the range of

secondary loadings seems quite difficult to specify on an a priori basis for all inventories, and we suspect that several of the minor secondary loadings that are likely to emerge in a given sample might fail to replicate on cross-validation. We found several instances of this phenomenon in our split-half analyses.

Second, our empirical illustration demonstrates limitations with "golden rules" as they are applied to the evaluation of model fit in a CFA context. Marsh et al. (2004) also raised important concerns about the limitations of universal "golden rules" for evaluating model fit and concluded that "interpretations of the degree of misspecification should ultimately have to be evaluated in relation to substantive and theoretical issues that are likely to be idiosyncratic to a particular study" (p. 340). They suggested that model fit "rules of thumb" are seductive but also pointed out that there is little evidence that they are appropriate for all contexts. One thing seems clear to us: Any omnibus personality inventory that shows adequate fit in CFA models by the criteria we selected as reflecting current conventions would mark quite an achievement.

In light of these and similar other results, some authors have suggested that exploratory methods should continue to be emphasized in research on personality trait structure (e.g., Goldberg & Velicer, 2006; Lee & Ashton, 2007). Indeed, although from a CFA perspective it appeared that none of the measures we examined had "acceptable" structures, a more nuanced picture emerged from EFA methods in which some measures consistently outperformed others. For instance, measures whose development was guided by EFA techniques such as the HEXACO, 6fpq, and NEO-PI-R tended to be relatively more amenable to factorial recovery than those measures whose authors were more explicitly dismissive of factor analysis in the initial scale development process (e.g., CPI; Gough & Bradley, 2002). Accordingly, we believe that EFA methods should continue to play an important role in personality science and that in many cases exploratory methods can be more informative than CFA methods for developing a better understanding of the structure of omnibus personality trait measures.

Moreover, several potentially useful methods have been developed for using EFA methods to test the structure of multiscale inventories that go beyond a simple "eyeball" test and that probably should not be described as *exploratory* in the loose sense of the word. We illustrated a few of these options such as our efforts to quantify the proportions of pattern coefficients above or below certain cutoffs or the computation of within-sample congruence coefficients. Likewise, we showed that computing congruence coefficients after rotating observed data to target matrices is an approach with considerable potential. In these analyses, we illustrated how researchers can use two different kinds of target matrices—one based on binary specifications and one based on data from previous samples. Although Procrustes rotation in general appears to be an appealing strategy, there did appear to be an advantage to using a previous sample as this method takes advantage of preexisting knowledge about cross-loadings. For example, when a previous matrix was used, our results were consistent with those of McCrae et al. (1996) in showing that, for the NEO-PI-R, most scales had strong congruence with a target matrix (congruence coefficients ranged from .93 to .97; Table 4, p. 561) even though the measure did not fit particularly well in a CFA framework, and these coefficients were considerably lower in our binary analyses. We suspect that substantial cross-loadings will occur for inventories that are valuable for predictive purposes (see, e.g., Ashton et al., 2009), and thus we recommend using target matrices from actual data when they are available rather than binary codes for Procrustes rotations. The broader point, however, is that our results are consistent with McCrae et al.'s suggestion that Procrustes rotations are a useful approach for quantifying the structural validity of omnibus personality trait measures.

There are also other approaches that can be used to further evaluate the structure of personality inventories in addition to the methods we used for our empirical illustrations. One approach is to attempt to explicitly model response style factors that can contribute to model misfit. For example, if a person has a generally favorable opinion of herself or himself, she or he might tend to respond to items with varying psychological content in a similar fashion. For example, to the extent she or he believes that both Openness and Extraversion are "good" and to the extent that she or he sees herself or himself as "good," she or he might be more likely to endorse characteristics representing either of these traits. Analyses of response style factors are particularly informative when both self- and other-report data are available so as to judge the correspondence between response style factors across different methods. For example, McCrae and Costa (2008) showed that modeling evaluative factors in a joint analysis of self- and other-report NEO-PI-3 data improved congruence coefficients among the five substantive factors. Specifying method factors for positively and negatively keyed items tends to improve the fit of many personality measures in confirmatory analyses as well (see, e.g., Quilty et al., 2006). However, an important and sometimes contentious issue is whether these response style and method factors are more than simple artifacts (e.g., Marsh, 1996), and future work is needed to more thoroughly evaluate whether such method factors have predictive utility.

Another promising method for evaluating test structure involves the use of exploratory structural equation modeling (ESEM) as articulated by Asparouhov and Muthén (2009; also see Marsh et al., 2009). In essence, ESEM allows users to freely estimate the extensive cross-loadings for personality indicators in the context of the general structural equation modeling framework or to use targeted matrices for their specification. One advantage of this technique is the ability to estimate latent variables for broad trait dimensions that

can be used for testing the correlates of personality or for evaluating measurement invariance. As a result, factor intercorrelations for theoretically orthogonal traits (e.g., those of the Big Five) are likely to be minimized, unlike in typical CFA contexts in which the factor correlations for measures operationalizing such systems are quite large, in part because of the myriad cross-loadings that often go unspecified (e.g., the cross-loading of the NEO-PI-R neuroticism impulsiveness facet on Conscientiousness; see Marsh et al., 2004). Artificially large higher order trait correlations will create problems for multivariate prediction research given the interpretational difficulties presented by correlated predictors (see Lynam, Hoyle, & Newman, 2006). Using ESEM, however, these issues are far less of a concern because the extensive cross-loadings are estimated, just as they are in EFA contexts. Thus, at a broad level, we suspect that this "hybrid" method might render contentions about the relative merits of EFA versus CFA less central and may instead allow researchers to focus on important substantive questions regarding personality structure and assessment.

Recommendations and Conclusions

In the service of a candid conversation about the role of factor analysis in testing the validity of personality trait instruments, we recommend that researchers judge the quality of a particular factor solution against findings from previous research on the inventory being evaluated or results for similar inventories when such specific information is not available. Indeed, previous studies might provide a more reasonable context for interpreting overall fit statistics for omnibus personality inventories rather than the rules of thumb widely used for covariance structure modeling applications. The overall model fit values reported here provide reasonable indications of what might be expected when evaluating omnibus inventories in reasonably sized samples.

We also echo Loevinger (1957) in asserting that internal structure should be regarded as just one element of construct validity among several others. In line with the importance of multiple kinds of validity evidence, we suggest that future research should thoughtfully combine the analysis of internal structure with investigations of criterion-related validity to a more substantial degree. Specifically, researchers should consider the implications of model modifications for theory and test and provide more information about what meaningful consequences any post hoc model modifications have for criterion-related validity. In many cases, we suspect that model modifications will have trivial implications for external validity and personality theory; however, exceptions to this rule are likely, and this would seem to be an important direction for future work. Likewise, we believe that criticisms of measures based solely on model misfit should be supplemented with evidence regarding the impact of suboptimal structure on other forms of validity. That is, we believe that there is a need to document that misspecifications have practical or substantive consequences beyond simply contributing to model misfit (Saris et al., 2009). This will help researchers determine whether simplifying assumptions about the structure of personality have more than trivial consequences. It is helpful for applied researchers to recall the maxim that models can be useful even if they simplify reality (Meehl, 1990).

With this maxim in mind, we further hope that these results generate pause in those critics of the structure of any one particular personality trait instrument as it appears that none of the most widely used personality inventories are beyond reproach using CFA to test internal structure. To be sure, we believe that it is fairly easy to identify apparent flaws in omnibus personality measures using CFA approaches. This is especially problematic when stringent CFA standards are applied to question the validity of new instruments, even when those same standards would tend to cast considerable doubt on the most widely used personality inventories that have shown considerable criterion-related validity (McCrae et al., 1996). Given that newer instruments, by definition, will not have shown extensive evidence of external validity, criticizing them purely on structural grounds in a CFA framework seems to us to represent a double standard that should be faced more squarely by researchers, reviewers, and editors.

In closing, we propose three summary recommendations for researchers interested in evaluating the internal structure of multidimensional personality inventories. First, researchers should utilize multiple factor analytic methods and report a range of exploratory and confirmatory analytic results. Second, researchers should avoid broad conventions to provide a "thumbs up or down" decision regarding overall model fit. Instead, test structure should be considered in the context of previously reported results and the substantive meaning of the parameters that contribute to overall misfit. Third, researchers should contextualize internal structure by considering its impact on criterion-related validity and its fidelity to a measure's underlying theoretical assumptions. Related to this point, researchers should formally evaluate the practical and conceptual consequences of model modifications used to achieve better fit.

Critics may suggest that each of these recommendations introduces subjective judgment into the process of validation research and may encumber progress in personality and psychological science more generally. To such critics, we respond that judgment has always been an "essential ingredient" in the research process (Cohen, 1990), and we believe that a wider context for judgments about the internal structure of personality measures will be more likely to facilitate scientific progress than to impede it. At the very least, such an approach might combat exaggerated concerns that the sky is falling on personality assessors.

Acknowledgments

We thank Lew Goldberg and Maureen Barckley for making the Eugene Springfield Community Sample data available for this

study. We also thank Michael Ashton for providing normative HEXACO data. Finally, we thank Emily Ansell, Dan Blonigen, Rachel Dinero, Mike Friedman, Lew Goldberg, Bob Krueger, Rich Lucas, Les Morey, Fred Oswald, Aaron Pincus, Rick Robins, Neal Schmitt, Seth Schwartz, and Keith Widaman for helpful comments on earlier drafts.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interests with respect to the authorship and/or publication of this article.

Financial Disclosure/Funding

The authors received no financial support for the research and/or authorship of this article.

Notes

1. Goodness-of-fit indices can be computed for exploratory solutions with maximum likelihood estimation given that both exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) can be understood in the context of the common factor model. However, these indices appear relatively infrequently in the applied literature. Moreover, these statistics are based on the least restrictive model for a given number of common factors, which is much less restrictive than typical CFA models tested for personality measures. Thus, greater permissiveness of EFA relative to CFA is because of both mathematical and practical differences between the methods.
2. Method factors can also be created by a subset of careless responders in a data set such as those individuals who endorse the same response for all items, regardless of their polarity (Schmitt & Stults, 1985).
3. Cloninger's Temperament and Character Inventory was administered to this sample, but it is not considered here because its internal structure in the Eugene Springfield Community Sample data was the subject of a previous article (Farmer & Goldberg, 2008). Notably, it did not fit the data well in a CFA framework.
4. The HEXACO college student development sample ($N = 1,681$) was used for these analyses (see Lee & Ashton, 2006). The GEOMIN output from these analyses are available on request.
5. All matrices are available from the first author on request.
6. The magnitudes of and variables involved in these modification indices are available from the first author on request.
7. The relatively poor performance of the California Psychological Inventory may be in part because of scales with overlapping items, a consequence of the emphasis its developers put on criterion-related, as opposed to structural, validity. The ability to achieve evidence of adequate Multidimensional Personality Questionnaire structure may have been limited in part to its somewhat low scale to factor ratio (especially for the four-factor model, which breaks Positive Emotionality into two components).
8. The 16PF and Hogan Personality Inventory (HPI) yielded factor solutions that were quite dissimilar to their theoretical models. Specifically, the 16PF Self-Control factor and the HPI Ambition factor were not recovered. Thus, we computed coefficients for Table 3 based on results from the other factors. Despite this permissive approach, these measures still did not perform well relative to other instruments.

References

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin, 103,* 411-423.

Ashton, M. C., & Lee, K. (2007). Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and Social Psychology Review, 11,* 150-166.

Ashton, M. C., Lee, K., Goldberg, L. R., & de Vries, R. E. (2009). Higher order factors of personality: Do they exist? *Personality and Social Psychology Review, 13,* 79-91.

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling, 16,* 397-438.

Barrett, P. (1986). Factor comparison: An examination of three methods. *Personality and Individual Differences, 7,* 327-340.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88,* 588-606.

Borkeneau, P., & Ostendorf, F. (1990). Comparing exploratory and confirmatory factor analysis: A study on the 5-factor model of personality. *Personality and Individual Differences, 11,* 515-524.

Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association, 71,* 791-799.

Brannick, M. T. (1995). Critical comments on applying covariance structure modeling. *Journal of Organizational Behavior, 16,* 201-213.

Breckler, S. J. (1990). Applications of covariance structure modeling in psychology: Cause for concern? *Psychological Bulletin, 107,* 260-273.

Briggs, S. R., & Cheek, J. M. (1986). The role of factor analysis in the development and evaluation of personality scales. *Journal of Personality, 54,* 106-148.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research.* New York: Guilford.

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research, 36,* 111-150.

Byrne, B. M. (2001). *Structural equation modeling with AMOS: Basic concepts, applications, and programming.* Mahwah, NJ: Lawrence Erlbaum.

Byrne, B. M. (2005). Factor analytic models: Viewing the structure of an assessment instrument from three perspectives. *Journal of Personality Assessment, 85,* 17-32.

Chan, W., Ho, R. M., Leung, K., Chan, D. K. S., & Yung, Y. F. (1999). An alternative method for evaluating congruence coefficients with Procrustes rotation: A bootstrap procedure. *Psychological Methods, 4*, 378-402.

Church, A. T., & Burke, P. J. (1994). Exploratory and confirmatory tests of the Big Five and Tellegen's three- and four-dimensional models. *Journal of Personality and Social Psychology, 66,* 93-114.

Cloninger, C. R. (2008). The psychobiological theory of temperament and character: Comment on Farmer and Goldberg. *Psychological Assessment, 20,* 292-299.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45,* 1304-1312.

Conn, S. R., & Rieke, M. L. (1994). *The 16PF fifth edition technical manual*. Champaign, IL: Institute for Personality and Ability Testing.

Costa, P. T., Jr., & 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.

Cudeck, R., & MacCallum, R. C. (2007). *Factor analysis at 100: Historical developments and future directions*. Mahwah, NJ: Lawrence Erlbaum.

Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The Satisfaction With Life Scale. *Journal of Personality Assessment, 49,* 71-75.

DiStefano, C., & Motl, R. W. (2009). Personality correlates of method effects due to negatively worded items on the Rosenberg Self-Esteem Scale. *Personality and Individual Differences, 46,* 309-313.

Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2006). The mini-IPIP scales: Tiny-yet effective measures of the Big Five factors of personality. *Psychological Assessment, 18,* 192-203.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4,* 272-299.

Farmer, R. F., & Goldberg, L. R. (2008). A psychometric evaluation of the revised Temperament and Character Inventory (TCI-R) and the TCI-140. *Psychological Assessment, 20,* 281-291.

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment, 7,* 286-299.

Gignac, G. E., Bates, T. C., & Jang, K. (2007). Implications relevant to CFA model misfit, reliability, and the five factor model as measured by the NEO-FFI. *Personality and Individual Differences*, *43*, 1051-1062.

Goldberg, L. R. (1993). The structure of personality traits: Vertical and horizontal aspects. In D. C. Funder, R. D. Parke, C. Tomlinson-Keasy, & K. Widaman (Eds.), *Studying lives through time: Personality and development* (pp. 169-188). Washington, DC: American Psychological Association.

Goldberg, L. R., & Velicer, W. F. (2006). Principles of exploratory factor analysis. In S. Strack (Ed.), *Differentiating normal and abnormal personality* (2nd ed., pp. 209-237). New York: Springer.

Gough, H. G., & Bradley, P. (1996). *CPI manual* (3rd ed.). Mountain View, CA: Consulting Psychologists Press.

Grucza, R. A., & Goldberg, L. R. (2007). The comparative validity of 11 modern personality inventories: Predictions of behavioral acts, informant reports, and clinical indicators. *Journal of Personality Assessment, 89,* 167-187.

Guadagnoli, E., & Velicer, W. F. (1991). A comparison of pattern matching indices. *Multivariate Behavioral Research, 26*, 323-343.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9,* 139-164.

Hoelzle, J. B., & Meyer, G. J. (2009). The invariant component structure of the Personality Assessment Inventory (PAI) full scales. *Journal of Personality Assessment*, *91*, 175-186.

Hofstee, W. K. B., de Raad, B., & Goldberg, L. R. (1992). Integration of the Big Five and circumplex approaches to trait structure. *Journal of Personality and Social Psychology, 63,* 146-163.

Hogan, R., & Hogan, J. (1995). *Hogan Personality Inventory manual* (2nd ed.). Tulsa, OK: Hogan Assessment Systems.

Hoyle, R. H., & Duvall, J. L. (2004). Determining the number of factors in exploratory and confirmatory factor analysis. In D. Kaplan (Ed.), *Handbook of quantitative methodology for the social sciences* (pp. 301-315). Thousand Oaks, CA: Sage.

Hu, L. T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3,* 424-453.

Jackson, D. N. (1971). The dynamics of structured personality tests: 1971. *Psychological Review, 78,* 229-248.

Jackson, D. N., Paunonen, S. V., & Tremblay, P. F. (2000). *Six Factor Personality Questionnaire*. Port Huron, MI: Sigma Assessment Systems.

Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The sources of four commonly reported cutoff criteria: What did they really say? *Organizational Research Methods, 9,* 202-220.

Lee, K., & Ashton, M. C. (2004). Psychometric properties of the HEXACO Personality Inventory. *Multivariate Behavioral Research, 39,* 329-358.

Lee, K., & Ashton, M. C. (2006). Further assessment of the HEXACO Personality Inventory: Two new facet scales and an observer report form. *Psychological Assessment, 18,* 182-191.

Lee, K., & Ashton, M. C. (2007). Factor analysis in personality research. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 424-443). New York: Guilford.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3*, 635-694.

Lorenzo-Seva, U., & ten Berge, J. M. F. (2006). Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology, 2,* 57-64.

Louks, J., Hayne, C., & Smith, J. (1989). Replicated factor structure of the Beck Depression Inventory. *Journal of Nervous and Mental Disease, 177,* 473-479.

Lykken, D. T. (1971). Multiple factor analysis and personality research. *Journal of Experimental Research in Personality, 5,* 161-170.

Lynam, D. R., Hoyle, R. H., & Newman, J. P. (2006). The perils of partialling: Cautionary tales from aggression and psychopathy. *Assessment*, *13*, 328-341.

MacCallum, R. (1986). Specification searches in covariance structure modeling. *Psychological Bulletin, 100*, 107-120.

Marsh, H. W. (1996). Positive and negative global self-esteem: A substantively meaningful distinction or artifactors? *Journal of Personality and Social Psychology, 70,* 810-819.

Marsh, H. W., Hau, K. T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indices and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling, 11,* 320-341.

Marsh, H. W., Muthen, B., Asparaouhov, T., Ludtke, O., Robitzsch, A., Morin, A. J. S., et al. (2009). *Exploratory structural equation modeling, integrating CFA and EFA: Application to students' evaluations of university teaching*. Unpublished manuscript.

Maydeu-Olivares, A., & Coffman, D. L. (2006). Random intercept item factor analysis. *Psychological Methods, 11,* 344-362.

McCrae, R. R., & Costa, P. T., Jr. (2008). Empirical and theoretical status of the five-factor model of personality traits. In G. Boyle, G. Matthews, & D. H. Saklofske (Eds.), *The SAGE handbook of personality theory and assessment* (pp. 273-294). Thousand Oaks, CA: Sage.

McCrae, R. R., Zonderman, A. B., Costa, P. T., Jr., Bond, M. H., & Paunonen, S. V. (1996). Evaluating replicability of factors in the revised NEO Personality Inventory: Confirmatory factor analysis versus Procrustes rotation. *Journal of Personality and Social Psychology, 70,* 552-566.

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry, 1,* 108-141.

Mulaik, S. A. (2006). *Foundations of factor analysis* (2nd ed.). New York: CRC Press.

Quilty, L. C., Oakman, J. M., & Risko, E. (2006). Correlates of the Rosenberg Self-Esteem Scale and method effects. *Structural Equation Modeling, 13,* 99-117.

Rosenberg, M. (1965). *Society and the adolescent self-image.* Princeton, NJ: Princeton University Press.

Russel, D. W. (2002). In search of underlying dimensions: The use (and abuse) of factor analysis in Personality and Social Psychology Bulletin. *Personality and Social Psychology Bulletin, 28*, 1629-1646.

Saris, W. E., Satorra, A., & van der Veld, W. M. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling, 16,* 561-582.

Schmitt, N., & Stults, D. M., (1985). Factors defined by negatively keyed items: The result of careless respondents? *Applied Psychological Measurement, 9,* 367-373.

Slocum-Gori, S. L., Zumbo, B. D., Michalos, A. C., & Diener, E. (2009). A note on the dimensionality of Quality of Life scales: An illustration with the Satisfaction With Life Scale (SWLS). *Social Indicators Research, 92,* 489-496.

Steger, M. F. (2006). An illustration of issues in factor extraction and identification of dimensionality in psychological assessment data. *Journal of Personality Assessment, 86,* 263-272.

Tafarodi, R. W., & Milne, A. B. (2002). Decomposing global self-esteem. J*ournal of Personality, 70,* 443-483.

Teel, C., & Verran, J. A. (1991). Factor comparison across studies. *Research in Nursing and Health, 14,* 67-72.

Tellegen, A. & Waller, N. G. (2008). Exploring personality through test construction: Development of the Multidimensional Personality Questionnaire. In G. J. Boyle, G. Matthews and D. H. Saklofske, (Eds.). *Handbook of personality theory and testing, II.* (pp. 261-292). London: Sage.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications.* Washington, DC: American Psychological Association.

Thurstone, L. L. (1947). *Multiple-factor analysis*. Chicago: University of Chicago Press.

Tucker, L. R. (1951). *A method for synthesis of factor analysis studies* (Personnel Research Section Report 984). Washington, DC: Department of the Army.

Vassend, O., & Skrondal, A. (1995). Factor analytic studies of the NEO Personality Inventory and the five-factor model: The problem of high structural complexity and conceptual indeterminacy. *Personality and Individual Differences, 19*, 135-147.

Vassend, O., & Skrondal, A**.** (1997). Validation of the NEO Personality Inventory and the five factor model. Can findings from exploratory and confirmatory factor analysis be reconciled? *European Journal of Personality*, *11*, 147-166.

Widaman, K. F. (2007). Common factors versus components: Principals and principles, errors and misconceptions. In R. Cudeck and R. C. MacCullum (Eds.). *Factor Analysis at 100: Historical Developments and Future Directions.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Wrigley, C., & Neuhaus, J. E. (1955). The matching of two sets of factors. *American Psychologist, 10,* 418-419.