Journal of Research in Personality 47 (2013) 599–608

Contents lists available at SciVerse ScienceDirect

Journal of Research in Personality

journal homepage: www.elsevier.com/locate/jrp

Simple measures and complex structures: Is it worth employing a more complex model of personality in Big Five inventories?

Anne Herrmann [a,*], Hans-Rüdiger Pfister [b]

[a] Kalaidos University of Applied Sciences Switzerland, Department of Economics and Management, Jungholzstrasse 43, 8050 Zurich, Switzerland
[b] Leuphana University of Lueneburg, Institute of Experimental Industrial Psychology, Wilschenbrucher Weg 84a, 21335 Lueneburg, Germany

ARTICLE INFO

Article history:
Available online 18 May 2013

Keywords:
Personality
Big Five structure
Confirmatory factor analysis
Exploratory Structural Equation Modeling
Construct validity
Multitrait–multimethod
NEO PI-R
16PF

ABSTRACT

The poor performance of five-factor personality inventories in confirmatory factor analyses (CFAs) prompted some to question their construct validity. Others doubted the CFA's suitability and suggested applying Exploratory Structural Equation Modeling (ESEM). The question arises as to what impact the application of either method has on the construct validity of personality inventories. We addressed this question by applying ESEM and CFA to construct better-fitting, though more complex models based on data from two questionnaires (NEO PI-R and 16PF). Generally, scores derived from either method did not differ substantially. When applying ESEM, convergent validity declined but discriminant validity improved. When applying CFA, convergent and discriminant validity decreased. We conclude that using current personality questionnaires that utilize a simple structure is appropriate.

1. Introduction

Researchers who investigate normal adult personality have reached a consensus on five broad factors, often called the 'Big Five' (Goldberg, 1990), and on their conceptual definitions (Digman, 1990; McCrae & Costa, 1999; Norman, 1963). These factors are known as Neuroticism, Extraversion, Openness, Agreeableness, and Conscientiousness, although other terms are used as well. This general consensus has allowed for cumulative research and meta-analyses of important aspects of the construct, including the development of personality over an individual's lifespan (Judge, Higgins, Thoresen, & Barrick, 1999; Terracciano, McCrae, & Costa, 2010), differences between groups (Goldberg, Sweeney, Merenda, & Hughes, 1998; Schmitt, Realo, Voracek, & Allik, 2008), the existence of a general factor of personality (Musek, 2007; van der Linden, te Nijenhuis, & Bakker, 2010), a prediction of external criteria (Grucza & Goldberg, 2007; Hurtz & Donovan, 2000), and many more. In research and practice, personality is predominantly assessed using self-report questionnaires. Many of these questionnaires contain items that contribute to one of many first-order scales that are combined to represent the Big Five factors.

The internal structure of personality, i.e., the assignment of subscales to the five factors, has commonly been examined using an exploratory factor analysis (EFA; Aluja, Rossier, Garcia, & Verardi, 2005; Cattell & Cattell, 1995; Costa & McCrae, 1992b). This assignment is extremely important because it forms the basis for obtaining scores for the higher-order personality factors. In general, a simple structure (Thurstone, 1947) where each first-order scale is uniquely assigned to only one of the Big Five factors is assumed to be appropriate.

As in many other research areas in which constructs are assessed using self-report questionnaires, CFAs were eventually applied to personality data. The results of these studies were largely discouraging. The CFA model fit indices frequently exceeded proposed cut-off values for acceptable model fits and, based on CFA standards, did not confirm the simple structure (Church & Burke, 1994; Hopwood & Donnellan, 2010; McCrae, Zonderman, Costa, Bond, & Paunonen, 1996; Vassend & Skrondal, 2011). Several cross loadings (i.e., links between first-order scales and factors other than the originally postulated higher-order personality factors) usually needed to be included in the model to achieve an acceptable fit. The more complex models, however, were difficult to interpret and often displayed less of a good fit in cross-validation samples (e.g., Church & Burke, 1994; Hopwood & Donnellan, 2010).

This has raised concerns if the currently proposed composition of the broad factors provides an adequate assessment of an individual's personality. These higher-order scores are commonly used in research studies and in practical applications of personality instruments. Thus, confidence is required regarding the suitability of the Big Five factors as a 'common language' for describing personality. Adding additional cross loadings as suggested by CFA also changes the meaning of the observed scores. Subsequently, one must

* Corresponding author.
E-mail address: anne.herrmann@kalaidos-fh.ch (A. Herrmann).

0092-6566/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jrp.2013.05.004

question how the construct validity of personality instruments is affected when subscales contribute to more than one broad factor.

In the present study we address these concerns in two ways: First, we determine the 'change of scores' which – in this examination – refers to a difference in the relative position of an individual within a sample on the trait continuum measured as the correlation between the original scores and scores obtained after incorporating the CFA cross loadings. Second, we examine the impact on the instruments' construct validity resulting from the modified models.

To complement our investigation and consider more recent trends in factor analysis, we also apply Exploratory Structural Equation Modeling (ESEM; Asparouhov & Muthen, 2009), a method that integrates CFA and EFA. ESEM is less restrictive than CFA as it does not constrain the non-target loadings to be zero. In difference to CFA, in ESEM a model can be specified only with regard to the number of factors. Further restrictions can be added and tested using chi-square difference tests. In difference to EFA, ESEM provides typical CFA parameters, such as standard errors and goodness of fit statistics as well as the possibility to test for measurement invariance between groups and across time (Asparouhov & Muthen, 2009). Due to these possibilities and advantages of ESEM, it has been promoted to be applied in the psychometric evaluation of psychological instruments (Marsh, Liem, Martin, Morin, & Nagengast, 2011).

We applied a CFA and ESEM to data from 620 respondents who completed two established personality questionnaires (the NEO PI-R and the 16PF questionnaire). Using two different sets of modification criteria to determine cross loadings when conducting the CFA, we generated two more complex models for each instrument. We computed scores based on these modified CFA models using two different approaches: (a) we applied the scoring rules for the instrument provided in the respective test manual but added the additional subscales, as identified in the CFA and (b) we used the factor scores obtained from the respective modified CFA model. The first approach mirrors current usage in research, in which manifest, rather than latent, Big Five scores are employed (Barrick & Mount, 1996; Grucza & Goldberg, 2007; Hurtz & Donovan, 2000; Salgado, 2003). The second approach uses scores that correspond more directly with the CFA models. With regard to the application of ESEM, we used the factor scores obtained from applying the method from both instruments.

To assess the relative score changes, we computed correlations between scores from the original model and the scores obtained from the CFA and ESEM models. The results of this analysis support a more nuanced discussion of the discrepancy between current personality theories and the more complex model of personality, as suggested by the CFA. Applying ESEM offers further insight into how Big Five scores based on a more recent factor-analytical method.

To determine the impact on the questionnaires' construct validity, we applied the multitrait–multimethod (MTMM) approach, which was developed by Campbell and Fiske (1959), to the original model as well as the models proposed by CFA and ESEM. A comparison of the MTMM results across the models showed the extent to which the relationships within and between the five factors of both instruments changed as one moved from a simple to a more complex structure, thus determining changes in the convergent and discriminant validity.

Previous studies have focused mainly on investigating the congruence between results obtained from the EFA and CFA of an instrument without examining the impact of the observed discrepancies on scale scores and construct validity beyond the internal structure (e.g., Aluja, Blanch, & Garcia, 2005; Borkenau & Ostendorf, 1990; McCrae et al., 1996). In other studies, CFAs were applied to several instruments, but it was not determined how the relationships between the constructs were affected by changes in the model proposed by the CFAs (e.g., Church & Burke, 1994; Hopwood & Donnellan, 2010). In our study, we address those gaps by determining how the scores of and the relationships between personality scales change when the internal structure is more complex, as suggested by CFA. As a result, we extend the examination of construct validity beyond the internal structure to focus on changes in the convergent and discriminant validity within and across the two instruments. The study thus follows a suggestion made, among others, by Hopwood and Donnellan (2010) that "there is a need to document that misspecifications have practical or substantive consequences beyond simply contributing to model misfit" (p. 343).

Considering the complexities and difficulties in identifying the correct model in CFA based on modification indices and other model assessment criteria (Fan & Sivo, 2007; MacCallum, Roznowski, & Necowitz, 1992), we do not aim at determining the "true" model of personality. Instead, we provide an empirical illustration, i.e., to demonstrate by way of example the impact that this added complexity would have on scores and construct validity. By also applying ESEM to both instruments, we shed light on how this more recent but increasingly used method may affect the resulting factor scores and subsequently the instruments' construct validity.

2. Method

2.1. Measures

The data from two hierarchical self-report personality instruments were used in this study:

(1) Cattell's 16 Personality Factor Questionnaire, 5th Edition (16PF, Conn & Rieke, 1994) consists of 185 items with a three-choice response format that measures 16 primary factors. The 15 non-cognitive factors are then combined into five factors, commonly called 'global factors'.
(2) The Revised NEO Personality Inventory (NEO-PI-R, Costa & McCrae, 1992b) comprises 240 items with a five-point Likert response format. It assesses 30 facets of personality that are used to compute five higher-level domain scores.

The 16PF and the NEO-PI-R differ in that the first-order level of personality is described with 15 and 30 scales, respectively. An alignment exists, however, between the second-order level, where there is a NEO domain counterpart for each 16PF global factor. The counterparts for both instruments are 16PF-Extraversion and NEO-Extraversion, 16PF-Anxiety and NEO-Neuroticism, 16PF-Self-Control and NEO-Conscientiousness, 16PF-Independence and NEO-Agreeableness and, finally, 16PF-Tough-Mindedness and NEO-Openness (Cattell & Mead, 2008). The last two pairs are defined in the opposite direction.

Different views exist on when to consider a psychometric questionnaire a "Big Five Instrument". We follow a definition by McCrae and John (1992): "The five-factor model of personality is a hierarchical organization of personality traits in terms of five basic dimensions" (p. 175) which applies to the NEO PI-R as well as the 16PF. These two Big Five instruments were included in this study because they differ profoundly in their development and in the approach to computing the second-order factors. This method safeguards against drawing conclusions about personality constructs that are actually a result of characteristics of a particular instrument. The 16PF questionnaire was developed based on empirical analyses. An EFA of the item parcels was carried out to identify the primary personality traits. These primary factors were subjected to a second-order EFA to extract five global factors (Cattell

& Cattell, 1995). Based on the size of the EFA factor loadings, a set of contributing primary factors and their respective weightings were selected in the computation of each global factor score. The global factor computation is thus data-driven with regard to the assignment of primary factors and their relative importance. This approach also resulted in multiple assignments; six of the 15 first-order factors contributed to two global factors. The authors of the NEO PI-R instrument reached a consensus on five factors and used those factors as a starting point to develop a hierarchical model of personality (Costa & McCrae, 1992b). Based on psychological literature and conceptual considerations, six facets were then determined for each factor that reflected relevant and diverse aspects of the respective higher-order construct. The five second-order factor scores are computed using unit-weighting. In other words, a simple sum score was obtained by adding up the scores of the contributing six facets. Unlike the 16PF, each facet contributed to one domain only.

2.2. Sample

The sample used in this study included 620 respondents and was a subset of the Eugene Springfield Community Sample (ESCS; see Grucza & Goldberg, 2007 for information on data collection procedures and further sample details). The original ESCS sample was slightly larger ($N = 857$ for the NEO PI-R and $N = 680$ for the 16PF). In our study, only the 620 participants who completed both instruments were included in the sample. Of these, 97% were Caucasian, 57% were female, and roughly half the sample had achieved at least a college degree. The age of the respondents ranged from 18 to 85 years old ($M = 52$, $SD = 13$).

Because both questionnaires were completed by the same sample, unknown characteristics of different samples can be ruled out as an explanation for any observed differences between instruments. This set-up also allows us to examine the construct validity across both instruments.

2.3. Analyses

2.3.1. CFAs and ESEM

CFAs were conducted using the software R (2012) and the package 'lavaan' (Rosseel, 2011), which have been shown to generate the same results as other software packages (Narayanan, 2012), to examine the second-order structure of the 16PF and the NEO PI-R. Because the data were non-normally distributed, we used a robust maximum likelihood estimation method that provided robust standard errors and Satorra–Bentler scaled test statistics (Satorra & Bentler, 2001). The original models were specified as follows: for each model, all loadings of the manifest variables on the five factors were assumed to be zero, except for the latent factors to which the manifest variable was assigned in the original model as specified in the test manual. The covariances between the latent variables were freely estimated in both models. Cattell used an oblique factor rotation when conducting the EFA of the 16PF during questionnaire development because it reflected his idea of interrelated personality factors (Cattell & Cattell, 1995). While the theoretical NEO PI-R model proposed orthogonal domains, the five domains displayed considerable intercorrelations which have been attributed to conceptual overlap of facets that may relate to more than one broad factor (Costa & McCrae, 1992a).

We used four fit indices that cover different aspects of model fits that were identified as particularly suitable for personality data, in which comparatively low target loadings and several secondary loadings were expected (Beauducel & Wittmann, 2005): (a) the Comparative Fit Index (CFI), an incremental fit index; (b) the Standardized Root Mean Square Residual (SRMR), an absolute fit; (c) the Root Mean Square Error of Approximation (RMSEA), an index that favors simple over more complex models; and (d) the Satorra–Bentler corrected χ^2 (SB-χ^2) test, a significance test used when data are distributed non-normally, as was the case in our study.

The purpose of our study was to gauge the impact that the cross loadings suggested by CFA had on Big Five scores. Hence, we decided to apply two different approaches to model modification in the analyses, resulting in two alternative CFA models per personality instrument. Both approaches reflected different ideas about what should guide modifications and what type of modifications were justifiable. The modification process for Model 1, the first alternative CFA model, was guided by the modification index (MI), which provides the researcher with a direct measure of the change in the model fit chi-square if the parameter was freed. Starting from the original model for both instruments, we computed MIs for successive models, each time freeing the factor loading or residual correlation between subscales with the highest MI until an acceptable model fit was obtained. There is a lively debate on the appropriateness of using general cut-off values for goodness-of-fit statistics (Marsh, Hau, & Wen, 2004). In the absence of hard-and-fast rules, we opted for the frequently applied cut-off values suggested by Hu and Bentler (1999): a CFI greater than .90, an SRMR less than .08, and an RMSEA less than .06.

For the second alternative CFA model, Model 2, the approach to model modification was guided by the intention to control for Type I as well as Type II errors (Saris, Satorra, & van der Veld, 2009). A Type I error is present if a parameter that is fixed to zero in the original model is classified as a misspecification and is therefore estimated in the revised model even though its population value is zero. A Type II error occurs when a parameter fixed to zero is not classified as a misspecification even though its population value is not zero (Hu & Bentler, 1998). We used the MI to identify paths to be freed, this time only releasing paths where the MI was greater than 10, thus applying a chi-square test with a significance level of .001 ($df = 1$). A large sample size, however, increases the likelihood of Type I errors for this value (Saris et al., 2009). Thus, we combined information provided by the MI with the Expected Parameter Change (EPC). This value indicates the size of a currently fixed parameter if it were to be freely estimated in a revised model. It is a standardized value that can be viewed as an effect size. There are no rules as to what minimum the EPC should take on to justify freeing the respective parameter. All loadings were included that fulfill a certain criterion, in this case those that indicated that at least 10% of the variance in the manifest variable was explained by the respective latent factor. Thus, we opted for a conservative cut-off value of .316 (absolute value). This value also lies between suggested values found in the literature, such as .30 (Kline, 1994) and .40 (Saris et al., 2009). We started from the original model, this time freeing the parameter with the highest MI and an EPC > .316 for each successive model until no further indices complied with the criteria outlined above. These rather conservative criteria were applied to avoid obtaining an over-fitted model that (a) is not replicable when fitted to another sample and (b) is not a parsimonious description of the relationships between variables (MacCallum et al., 1992). For Model 2, the model modifications were restricted to releasing paths between indicators and latent variables. Error terms between manifest variables were not included because their conceptual meaning had been questioned (Gerbing & Anderson, 1984).

ESEM was conducted using the software Mplus 6.11 (Muthén & Muthén, 2012) and applying an oblique geomin rotation, thus allowing the factors to covary. The same model fit indices as for the CFA were computed.

2.3.2. Correlations between scores from the original model and the CFA and ESEM models

For the first set of scores based on the modified CFA models (M1m and M2m), we applied the scoring rules for the instrument

provided in the respective test manual but added the additional subscales, as identified in the CFA. Thus, the modified NEO PI-R domain scores were obtained as a unit-weighted sum of the raw scores of the six original facets and the additional facets. The modified 16PF scores were computed as a weighted sum of the original and additional primary factors identified in the CFA. We applied the average weighting of the original primary factors for each global factor to the additional subscales, thus neither downplaying nor overestimating their impact. Research has shown that weighting only produces minor relative changes in scores compared to unit weighting, especially under conditions where the number of components is high, where these components are correlated and where their weights vary only slightly (Bobko, Roth, & Buster, 2007). All three conditions apply to the components contributing to the global factors of the three 16PF models. Therefore, we could rule out that the weighting unduly affects correlations between scores. For the second set of scores based directly on the CFA results (M1c and M2c), the factor scores of the respective modified CFA model for both instruments were calculated. In lavaan, these scores are estimated based on a regression method referred to as 'modal posterior estimator'. Correlations between the original model with the ESEM model are based on the ESEM factor scores (EM) of the respective instrument.

Two sets of correlations between scores from the original and the modified models for the factors of both instruments were computed: (a) Pearson correlation coefficients to measure the strength of the linear relationship between scores from the original and the modified models and (b) Spearman correlation coefficients to quantify the change in rank order, thus determining the concordance of the ordering of individuals on each broad domain between the original and the more complex models.

2.3.3. MTMM

We used the multitrait–multimethod matrix (MTMM) developed by Campbell and Fiske (1959) as a framework to compare the level of convergent and discriminant validity of both instruments across the original scores and the scores obtained from ESEM as well as from both CFA modified models based on the two different approaches to score computation. Thus, six MTMM matrices were computed. Convergent validity is confirmed when high correlations are observed for corresponding scales, i.e., for scales that measure the same constructs across both instruments (monotrait–heteromethod, MTHM). Discriminant validity reflects the idea that traits that are not conceptually related should display considerably lower correlations than the ones between corresponding traits. Discriminant validity is supported when the non-diagonal intercorrelation coefficients within one method (heterotrait–monomethod, HTMM) are low and the non-diagonal intercorrelation coefficients between the traits of the two methods (heterotrait–heteromethod, HTHM) are even lower (Campbell & Fiske, 1959). Further support of construct validity is provided when the pattern of correlations between traits is similar for both methods. To compare the evidence of construct validity of the original model with the modified models, we calculated the means for each set of coefficients constituting different aspects of convergent and discriminant validity for each model separately. For this purpose, we used Fisher's transformation because it has been shown to be the preferable procedure when averaging correlations (Silver & Dunlap, 1987).

3. Results

3.1. CFAs and ESEM

The original simple structure models underlying both instruments exhibited an unacceptable model fit when conducting CFA (NEO PI-R: SB-χ^2 = 3493.44, df = 395, p < .001, SRMR = .13, RMSEA = .11, CFI = .61; 16PF: SB-χ^2 = 669.94, df = 74, p < .001, SRMR = .08, RMSEA = .11, CFI = .76). Altogether, 42 modifications (29 released paths, 13 residual covariances) to the NEO PI-R were required to obtain an acceptable model fit for Model 1 (SB-χ^2 = 1116.85, df = 353, p < .001, SRMR = .06, RMSEA = .06, CFI = .90). For the 16PF, fewer modifications (six released paths, six residual covariances) needed to be included until an acceptable model fit was achieved for Model 1 (SB-χ^2 = 197.98, df = 62, p < .001, SRMR = .04, RMSEA = .06, CFI = .95). When applying the more conservative criteria (MI > 10 and EPC > .316) to derive Model 2, 12 and five paths were added to the NEO PI-R and the 16PF, respectively, until no fixed parameter fulfilled the a priori criteria for being freely estimated. Neither of the final two models achieved an acceptable model fit (NEO PI-R: SB-χ^2 = 2167.78, df = 383, p < .001, SRMR = .09, RMSEA = .09, CFI = .77; 16PF: SB-χ^2 = 391.29, df = 69, p < .001, SRMR = .05, RMSEA = .09, CFI = .87).

In Tables 1 and 2, we provide an overview of the contributing subscales for the original and the two CFA modified models of both instruments. Many, but not all, of the subscales added to the domains in Models 1 and 2 were logical. For example, it is plausible to assign Dominance to the 16PF factor of Extraversion. It is less intuitive, however, to know how Abstractedness is related to the 16PF factor of Anxiety. Similarly, it seems reasonable to add Warmth and Positive Emotion to the NEO domain of Agreeableness. The negative link between Aesthetics and the NEO domain of Extraversion, however, is hard to explain conceptually.

The application of ESEM provided better model fit due to the less restrictive assumptions (NEO PI-R: SB-χ^2 = 1231.49, df = 295, p < .001, SRMR = .03, RMSEA = .07, CFI = .90; 16PF: SB-χ^2 = 197.61, df = 40, p < .001, SRMR = .03, RMSEA = .08, CFI = .94). For the NEO PI-R, the facets displayed substantial loadings on their respective domain. Only very few higher non-target loadings were observed. For the 16PF, the subscales displayed substantial loadings on their respective factor. Only for Independence a less clear pattern emerged and some differences with regard to the assigned subscales according to the test manual were found.

3.2. Correlations between scores from the original model and the CFA and ESEM models

Pearson correlation coefficients between counterparts of the Big Five scores based on the original model and the CFA- and

Table 1
Overview of contributing primary factors for the 16PF models.

16PF global factor	Contributing 16PF primary factors in the original model	Additional primary factors in Model 1	Additional primary factors in Model 2
Anxiety	Vigilance (L), Apprehension (O), Tension (Q4), Emotional Stability (−C)	–	M, E
Extraversion	Warmth (A), Liveliness (F), Social Boldness (H), Privateness (−N), Self-Reliance (−Q2)	E	E
Tough-Mindedness	Warmth (−A), Sensitivity (−I), Abstractedness (−M), Openness to Change (−Q1)	Q4, F, L	C
Independence	Dominance (E), Vigilance (L), Social Boldness (H), Openness to Change (Q1)	−O	−O
Self-Control	Rule-Consciousness (G), Abstractedness (−M), Perfectionism (Q3), Liveliness (−F)	N	–

Note: '−' Indicates a reversed loading of the primary factor onto the global factor.

Table 2
Overview of contributing facets for the NEO PI-R models.

NEO PI-R Domain	Contributing NEO PI-R facets in the original model	Additional facets in Model 1	Additional facets in Model 2
Neuroticism	Anxiety (n1); Angry Hostility (n2); Depression (n3); Self Consciousness (n4); Impulsiveness (n5); Vulnerability (n6)	−c1, a5, −a1, −c5, −c6, o3	−c1, a5, −e3
Extraversion	Warmth (e1); Gregariousness (e2); Assertiveness (e3); Activity (e4); Excitement Seeking (e5); Positive Emotions (e6)	−a2, n5, c4, a3, a1, −o2, −c6, o3	a3, a1, c4
Openness	Fantasy (o1); Aesthetics (o2); Feelings (o3); Actions (o4); Ideas (o5); Values (o6)	a6	–
Agreeableness	Trust (a1); Straightforwardness (a2); Altruism (a3); Compliance (a4); Modesty (a5); Tender Mindedness (a6)	e1, e6, −n2, e2, c3, −n3, o3	e1, −n2, e6, e2
Conscientiousness	Competence (c1); Order (c2); Dutifulness (c3); Achievement Striving (c4); Self-Discipline (c5); Deliberation (c6)	e4, e3, −o1, −a1, −n5, a3, −n6	−n5, e4

Note: '−' Indicates a reversed loading of the facet onto the domain.

Table 3
Pearson correlation coefficients of the 16PF and NEO PI-R factors across models.

	OM–M1m		OM–M2m		OM–RV	OM–M1c	OM–M2c	OM–EM
	r	r_r	*r*	r_r	r_{rv}	*r*	*r*	*r*
NEO PI-R[a]								
Neu.	.95	.90	.95	.94	.68	.95	.97	.98
Ext.	.93	.84	.95	.92	.63	.80	.78	.87
Open.	.99	.98	n.a.	n.a.	.67	.96	.98	.98
Agree.	.82	.80	.88	.88	.61	.92	.92	.90
Cons.	.93	.83	.97	.94	.60	.97	.98	.97
16PF[a]								
Anx.	n.a.	n.a.	.90	.87	.59	−.93[b]	−.89[b]	−.93[b]
Ext.	.97	.97	.97	.97	.73	.90	.96	.97
T-M.	.85	.85	.95	.95	.65	−.70[b]	−.72[b]	−.79[b]
Ind.	.95	.94	.95	.93	.60	.60	.52	.62
S-C.	.94	.94	n.a.	n.a.	.65	.79	.89	.82

Note: $N = 620$; Neu. = Neuroticism, Ext. = Extraversion, Open. = Openness to Experience, Agree. = Agreeableness, Conc. = Conscientiousness, Anx. = Anxiety, T.-M. = Tough-Mindedness, Ind. = Independence, S-C. = Self-Control, OM = Original model, M1m/M2m = Model 1/Model 2 scores computed based on scoring rules from respective test manual and additional scales included as suggested by CFA, M1c/M2c = Model 1/Model 2 CFA factor scores, EM = Exploratory Structural Equation Modeling factor scores, RV = Random Variable Model, *r* = Pearson correlation coefficient, r_r = Pearson correlation coefficient with random variables added to M1 and M2, r_{rv} = Pearson correlation coefficient with maximum number of random variables added to the model, n.a. = not applicable as scores were the same for the two models.

[a] All correlations $p < .001$.

[b] Factor scores are reversed.

ESEM-based scores for the NEO PI-R and the 16PF are shown in Table 3. The Spearman coefficients were almost identical to the Pearson coefficients (maximum difference .03). As such high similarity was found, and in order to save journal space, the Spearman coefficients are not reported.

The Pearson coefficients of the original model scores with the modified CFA scores which were computed as instructed by the respective test manual but with the additional subscales suggested by CFA (M1m and M2m) were fairly high. They ranged from .82 to .99 for the NEO PI-R and from .85 to .97 for the 16PF. Only three coefficients were below .90. The Pearson coefficients of the original model scores with the CFA factors scores obtained from the modified models (M1c and M2c) were also fairly high for the NEO PI-R (.78–.98). However, a reduced agreement was found for the 16PF (.52–.96), with particularly low coefficients for Tough-Mindedness and Independence. A similar pattern emerged for the ESEM factor scores (EM): Fairly high Pearson coefficients with the original model scores were obtained for the NEO PI-R (.87–.98). The agreement for the 16PF was in general lower (.62–.97). The lowest coefficient was obtained for Independence, the factor which also displayed the least clear pattern of subscale loadings in the ESEM solution.

When comparing the agreement of the original scores with both sets of CFA-based modified scores (M1c/M2c versus M1m/M2m), the M1m/M2m scores displayed a higher agreement with the original Big Five scores. The M1m and M2m scores are computed following the instructions in the manual, albeit with some subscales added as suggested by CFA. Thus, the score obtained from the respective modified model M1m and M2m still contains the four to six subscales and applies the same weighting to these subscales as in the original model. The correlation coefficient between scores from the original and the modified CFA model M1m and M2m is therefore always in large part a correlation with itself. The M1c and M2c scores of the modified models also share the four to six subscales with the original model. However, the weighting of these subscales in the score computation was based on the CFA factor loadings and hence may differ from what was applied in the original model scores, thus offering one explanation for the slightly lower agreement.

To gauge the impact of this shared variance of scores between the original and the modified CFA models M1m and M2m, we generated a set of random subscales with scores for each respondent. These new variables were specified to have means and standard deviations similar to the subscales of the two instruments and zero-correlations with each other and with the original subscales. Using scores from these random variables, we computed two matching sets of alternative broad factor scores for each individual based on the two alternative models for both questionnaires.

For the first set of alternative scores, we added the same number of random variables to the computation of each broad factor score as was added to obtain scores for the two modified models of each questionnaire. For example, based on the CFA, eight additional facets were assigned to the factor of Extraversion in the NEO PI-R Model 1. Thus, we added eight random variables when computing the NEO PI-R Extraversion scores for the random-variable Model 1. The correlation coefficients, r_r, between the original model scores and the scores for random-variable Model 1 and random-variable Model 2 were only marginally smaller or sometimes equal to the coefficients obtained when adding scales based on a CFA of the original data and applying the scoring rules in the respective test manual (see Table 3). Therefore, adding the same number of zero-correlated random variables to the original model creates just as much relative change in the original scores as does the addition of scales identified by the CFA to the broad construct.

For the second set of alternative scores, we added the maximum number of random variables to the computation of each broad factor score, considering the overall number of narrow scales in each instrument. Thus, 24 random variables were added when computing each of the five NEO scores, and 10 or 11 random variables were added when computing each of the five 16PF scores for these random-variable models. We then computed Pearson correlation coefficients, r_{rv}, between the original scores and the scores obtained from this random-variable procedure for both models across both questionnaires. These correlation coefficients, shown in Table 3, were considerably smaller, ranging from .60 to .68 for the NEO PI-R and from .59 to .73 for the 16PF.

Table 4
Multitrait–multimethod correlation matrix of the original model and for the ESEM model.

	NEO PI-R					16PF				
	Neu.	Ext.	Open.	Agree.	Cons.	Anx.	Ext.	T-M.	Ind.	S-C.
NEO PI-R										
Neu.	(.94/.95)	**−.28**	**−.16**	**.04**	**−.32**	*.56*	*.00*	*−.14*	*−.52*	*−.07*
Ext.	**−.30**	(.91/.94)	**.45**	**.00**	**.17**	*−.10*	*.59*	*.61*	*.33*	*.02*
Open.	**−.05**	**.33**	(.92/.93)	**−.17**	**.00**	*.02*	*.05*	*.40*	*.22*	*−.49*
Agree.	**−.21**	**.05**	**.04**	(.90/.92)	**−.03**	*−.32*	*.03*	*.08*	*−.53*	*.23*
Cons.	**−.44**	**.20**	**−.13**	**.14**	(.91/.93)	*−.04*	*−.05*	*.01*	*.26*	*.50*
16PF										
Anx.	*.68*	*−.31*	*−.14*	*−.24*	*−.21*	(.87/.87)	**−.05**	**−.12**	**−.13**	**−.15**
Ext.	*−.09*	*.66*	*.22*	*.17*	*−.05*	**−.29**	(.91/.85)	**.57**	**.16**	**.15**
T−M.	*−.05*	*−.23*	*−.66*	*−.10*	*.23*	**.04**	**−.41**	(.85/.90)	**.22**	**−.11**
Ind.	*−.14*	*.47*	*.34*	*−.35*	*.12*	**−.08**	**.38**	**−.38**	(.84/.89)	**−.03**
S−C.	*−.11*	*−.12*	*−.44*	*.24*	*.57*	**−.04**	**−.20**	**.49**	**−.22**	(.86/.88)

Note: $N = 620$; Neu. = Neuroticism, Ext. = Extraversion, Open. = Openness to Experience, Agree. = Agreeableness, Conc. = Conscientiousness, Anx. = Anxiety, T.-M. = Tough-Mindedness, Ind. = Independence, S-C. = Self-Control. Reliability coefficients are in parentheses (OM/EM); the monotrait-heteromethod correlations are underscored; the triangular heterotrait–monomethod matrices are in boldface; the square heterotrait–heteromethod matrices are in italics. Coefficients displayed in the lower-left triangle are based on the original model with scores computed based on scoring rules from the respective test manual. Coefficients displayed in the upper-right triangle are based on ESEM factor scores; Tough-Mindedness is reversed.
All correlations (absolute values) > .12 are $p < .001$, .09–.11 are $p < .01$, .06–.08 are $p < .05$, < .06 are *n.s.*

Table 5
Multitrait–multimethod correlation matrix of Model 1.

	NEO PI-R					16PF				
	Neu.	Ext.	Open.	Agree.	Cons.	Anx.	Ext.	T-M.	Ind.	S-C.
NEO PI-R										
Neu.	(.95/.95)	**−.43**	**−.20**	**.03**	**−.28**	*−.71*	*−.20*	*.18*	*−.33*	*−.19*
Ext.	**−.17**	(.90/.95)	**.54**	**−.62**	**−.06**	*.26*	*.34*	*−.01*	*.35*	*−.24*
Open.	**.04**	**.27**	(.92/.93)	**−.13**	**−.13**	*.09*	*.22*	*.36*	*.18*	*−.53*
Agree.	**−.48**	**.45**	**.28**	(.94/.95)	**.20**	*.09*	*.15*	*.28*	*−.47*	*.25*
Cons.	**−.66**	**.24**	**−.16**	**.31**	(.93/.93)	*.29*	*.05*	*−.22*	*.09*	*.43*
16PF										
Anx.	*.64*	*−.21*	*−.14*	*−.46*	*−.29*	(.87/.88)	**.42**	**−.12**	**.19**	**.45**
Ext.	*−.07*	*.65*	*.24*	*.36*	*.10*	**−.28**	(.91/.92)	**.56**	**−.45**	**.08**
T−M.	*.07*	*−.10*	*−.60*	*−.33*	*.12*	**.38**	**−.27**	(.88/.84)	**−.67**	**−.45**
Ind.	*−.28*	*.42*	*.28*	*−.01*	*.27*	**−.31**	**.53**	**−.17**	(.86/.85)	**−.18**
S−C.	*−.20*	*−.27*	*−.45*	*.04*	*.45*	**.04**	**−.43**	**.41**	**−.27**	(.87/.86)

Note: $N = 620$; Neu. = Neuroticism, Ext. = Extraversion, Open. = Openness to Experience, Agree. = Agreeableness, Conc. = Conscientiousness, Anx. = Anxiety, T.-M. = Tough-Mindedness, Ind. = Independence, S-C. = Self-Control. Reliability coefficients are in parentheses (M1m/M1c); the monotrait-heteromethod correlations are underscored; the triangular heterotrait–monomethod matrices are in boldface; the square heterotrait–heteromethod matrices are in italics. Coefficients displayed in the lower-left triangle are based on scores computed based on scoring rules from the respective test manual and additional scales included as suggested by CFA. Coefficients displayed in the upper-right triangle are based on CFA factor scores; Anxiety and Tough-Mindedness are reversed.
All correlations (absolute values) > .12 are $p < .001$, .09–.11 are $p < .01$, .06–.08 are $p < .05$, < .06 are n.s.

3.3. MTMM

The results of the MTMM-analyses are shown in Tables 4–6. Overall, convergent validity was supported for the original model (see lower-left triangle in Table 4): four of the five MTHM coefficients are considerable larger than all heterotrait coefficients. Only the relationship between NEO PI-R Agreeableness and 16PF Independence is smaller than two of the HTMM coefficients. Furthermore, the discriminant validity of the instruments is supported because the HTMM coefficients for both instruments are generally smaller than the MTHM coefficients and larger than the HTHM coefficients.

The pattern of the four MTMM matrices obtained for the two modified CFA models across both approaches to score computation is less clear (see Table 5 and 6). Evidence for convergent validity is less convincing because the correlation coefficients between the 16PF and NEO PI-R counterparts are consistently lower. Furthermore, evidence for the discriminant validity for these four MTMM matrices is weak as indicated by high correlations between conceptually unrelated factors across all four MTMM matrices based on the modified CFA model scores. Out of 80 HTMM and HTHM coefficients, 19 and 18 coefficients for Model 1 and Model 2, respectively, exceed an absolute value of .40.

And examination of the MTMM matrix based on the ESEM (see upper-right triangle in Table 4) showed that the convergent validity of the ESEM model was supported because the MTHM coefficients were of consistently high magnitude. More noteworthy however was the discriminant validity of the instruments assessed by the HTMM and the HTHM coefficients: Considerably lower correlations between conceptually unrelated Big Five factors based on the ESEM scores were obtained than based on the original model, particularly for the 16PF.

No absolute rules are available as to what can be considered sufficient evidence of construct validity based on MTMM results (Bagozzi & Yi, 1991). Instead, the pattern of correlation coefficients should be judged to assess the instrument's construct validity. To

Table 6
Multitrait–multimethod correlation matrix of Model 2.

	NEO PI-R					16PF				
	Neu.	Ext.	Open.	Agree.	Cons.	Anx.	Ext.	T-M.	Ind.	S-C.
NEO PI-R										
Neu.	(.94/.94)	**−.18**	**−.11**	**−.18**	**−.49**	−.53	*−.17*	*.31*	*−.42*	*.03*
Ext.	**−.51**	(.92/.93)	**.44**	**−.69**	**−.07**	*.05*	.40	*−.03*	*.27*	*−.29*
Open.	**−.11**	**.31**	(.92/.93)	**−.07**	**−.20**	*.22*	*.31*	.41	*−.04*	*−.57*
Agree.	**−.29**	**.59**	**.16**	(.93/.94)	**.25**	*.39*	*.11*	*.28*	−.42	*.25*
Cons.	**−.57**	**.43**	**−.09**	**.20**	(.92/.92)	*.15*	*.00*	*−.31*	*.18*	.44
16PF										
Anx.	.52	*−.24*	*.07*	*−.48*	*−.28*	(.88/.88)	**.61**	**.24**	**−.11**	**−.08**
Ext.	*−.19*	.63	*.23*	*.35*	*.04*	**−.10**	(.91/.91)	**.56**	**−.39**	**−.07**
T-M.	*−.22*	*−.08*	−.60	*−.08*	*.30*	**−.38**	**−.30**	(.85/.81)	**−.84**	**−.31**
Ind.	*−.45*	*.41*	*.32*	−.11	*.20*	**.04**	**.53**	**−.20**	(.86/.84)	**−.18**
S-C.	*−.09*	*.00*	*−.44*	*.14*	.51	**−.27**	**−.20**	**.51**	**−.21**	(.86/.86)

Note: N = 620; Neu. = Neuroticism, Ext. = Extraversion, Open. = Openness to Experience, Agree. = Agreeableness, Conc. = Conscientiousness, Anx. = Anxiety, T.-M. = Tough-Mindedness, Ind. = Independence, S-C. = Self-Control. Reliability coefficients are in parentheses (M2m/M2c); the monotrait-heteromethod correlations are underscored; the triangular heterotrait–monomethod matrices are in boldface; the square heterotrait–heteromethod matrices are in italics. Coefficients displayed in the lower-left triangle are based on scores computed based on scoring rules from the respective test manual and additional scales included as suggested by CFA. Coefficients displayed in the upper-right triangle are based on CFA factor scores; Anxiety and Tough-Mindedness are reversed.
All correlations (absolute values) > .12 are $p < .001$, .09–.11 are $p < .01$, .06–.08 are $p < .05$, <.06 are n.s.

judge whether the pattern of one model provides a stronger support of construct validity than another, we computed mean values for the five MTMM matrices separately using Fisher's transformation.

We found consistently weaker support of construct validity in the four MTMM matrices based on the modified CFA models in comparison to the original model. First, the mean values of the MTHM matrices are considerably smaller in the modified CFA models (.50, .49, .49, and .44 for M1m, M1c, M2m, and M2c respectively, compared to .59 in the original model), indicating a decline in convergent validity for the modified models. Second, a mean increase in the HTMM matrices of the modified CFA models was observed for the 16PF (.31, .37, .28, and .38 for M1m, M1c, M2m, and M2c respectively, compared to .26 in the original model), and more pronounced for the NEO PI-R (.32, .27, .34, and .28 for M1m, M1c, M2m, and M2c respectively, compared to .19 in the original model). The reduced differentiation between non-matching traits is caused by several subscales that now contribute to more than one factor, creating not only a conceptual overlap but also shared variance that leads to increased correlations among broad domains. Finally, there is also a mean increase for the HTHM matrices, albeit only marginal (.24, .21, .23, and .22 for M1m, M1c, M2m, and M2c respectively, compared to .20 for the original model). Overall, a considerable decline in convergent and discriminant validity compared to the original was obtained for all modified CFA models.

The results based on ESEM display a less consistent pattern and are therefore discussed separately. Compared to the original model, the convergent validity of the ESEM model was slightly reduced as indicated by a mean value of .52 for the MTHM matrix. In fact, four of the five coefficients based on ESEM factor scores were considerably smaller than in the original model. Interestingly, the Big Five factor Agreeableness/Independence, which typically displays the least agreement across both instruments, was found to be more similar when using ESEM factor scores (MTHM correlation coefficient of −.53, compared to −.35 in the original model). Particularly remarkable however is the improved discriminant validity of the instruments when using ESEM factor scores. A mean decrease in the HTMM matrices of the ESEM model was observed for the NEO PI-R (.17, compared to .19 in the original model), and more pronounced for the 16PF (.18, compared to .26 in the original model). A slight mean decrease was also observed for the HTHM matrix (.19, compared to .20 in the original model).

4. Discussion

CFA and ESEM were applied to two personality instruments based on the Big Five framework to determine the impact the factor structure suggested by these factor-analytical methods had on relative scores as well as on the construct validity of the NEO PI-R and the 16PF. MTMM analyses based on the Big Five scores of the CFA models revealed a considerable decrease in the convergent and discriminant validity of the questionnaires. Results based on ESEM were more promising in that the discriminant validity was improved in comparison to the original model. However, with the exception of Agreeableness/Independence, a considerable decrease in the convergent validity was observed.

The results – particularly those based on CFA models – highlight some important issues.

Several additional links between subscales and factors were suggested by CFA, indicating that the imposed simple internal structure may not be an adequate description of the construct personality. Introducing these additional links may indeed result in a model that better reflects the internal structure of personality. It has been argued before that the five factors are not as distinct as often suggested. In fact, even Costa and McCrae (1992b) acknowledge that some secondary loadings are "appropriate and meaningful" (p. 45), such as a high negative loading of the Neuroticism facet Angry Hostility on the domain Agreeableness. In the interest of simplicity and interpretability a decision was made to assign each subscale to one factor only and to exclude any additional relationships with other factors.

However, retaining the simple structure seems to be not only advisable in order to ensure the measures' interpretability. The present study shows that deciding against a more complex structure also avoided a negative impact on their convergent and discriminant validity. While introducing additional links in the models has led to an increase of internal validity by better reflecting the complex relationships between subscales and higher-order factors, this improvement was achieved at the expense of the instruments' convergent and discriminant validity which is not desirable. First, a decline in convergent validity resulted in a decreased consensus on the five broad personality factors. This impedes comparisons of research findings on personality conducted using different measures and will make it more difficult to combine them in meta-analyses. Thus, the more complex structure

jeopardizes the benefit of a five-factor framework. Second, a decline in discriminant validity as indicated by higher intercorrelations showed that the broad factors are conceptually less differentiated and hence might be less useful in applied settings. This tradeoff between the instrument's capability to adequately represent the complex internal structure of personality while preserving its convergent and discriminant validity cannot easily be resolved.

Furthermore, while there are good theoretical reasons to question the proposed simple structure, CFA should not be the method of choice to determine a more appropriate representation of the internal structure of personality. First, different assignments of scales to factors were obtained depending on the modification criteria and cut-off criteria applied in CFA, especially for the 16PF. Second, some of the additional links identified in CFA may not reflect conceptual relationships but are method artifacts, due to response styles such as social desirability (Ziegler & Buehner, 2009) or particular item content, such as negatively phrased items (Biderman, Nguyen, Cunningham, & Ghorbani, 2011). While these effects were not examined in the present study, it is important to remember that they may provide an explanation for some of the relationships found between subscales.

A high agreement between the original scores and the modified scores computed following the respective test manual (M1m and M2m) were obtained. Including an additional path in a CFA model when the respective subscale displays a high loading on a factor results in a small relative change in an individual's score. This is because very little additional variance is added. At the same time, the conceptual benefit is questionable because the constructs reflected by these modified composite scores become increasingly complex and less distinct with respect to their conceptual meaning.

Furthermore, by adding the same number of random scales as had been performed in the modified models, the magnitude of relative change was approximately the same as what had been obtained when adding scales suggested by the CFA, with the exception of Extraversion and Conscientiousness of the NEO PI-R. As the additional subscales were specified to be unrelated in this simulation, they quantify the maximum relative change that may occur in such an instance, regardless of which subscales may be assigned. This is particularly informative because the assignment of subscales to factors based on the CFA has been shown to depend on the decision criteria applied during the process of model modification. In the second simulation, the maximum number of subscales was added to each broad domain. It shows the maximum relative score change if one were to add all remaining subscales to each factor. While this may not present a realistic scenario, it offers a benchmark against which the observed differences between the original and the modified model can be judged.

A reduced agreement between the original scores with the CFA factor scores based on the modified CFA models (M1c and M2c) and ESEM scores were obtained, particularly for the 16PF. The fact that the original Big Five scores based on conventional scoring yield different results from applying CFA and ESEM factor scores has important implications for research and practice using personality questionnaires. First, a research study may yield different results depending on how the Big Five scores were obtained. Second, scores based on the modified CFA models and ESEM models are conceptually different constructs because their conceptual meaning is determined by the specific combination of contributing subscales. As such, potentially different findings between studies are not only likely but also plausible as analyses will be based on personality factors that do not share the same conceptual meaning. Third, regarding the applicability of research findings based on CFA and EFA factor scores, caution needs to be exercised as these results may not be directly transferable to practical applications where conventional scoring is used.

Personality questionnaires have repeatedly exhibited good criterion-related validity (e.g., Grucza & Goldberg, 2007; Hurtz & Donovan, 2000). The poor support of their internal structure has raised the question of how these measures can predict external criteria. However, the simple structure may in fact be beneficial for the measures' predictive capabilities. Several studies have demonstrated that broader domains reduce the predictive power of personality (Dudley, Orvis, Lebiecki, & Cortina, 2006; Tett, Steele, & Beauregard, 2003). In addition, from a conceptual and practical viewpoint, using these more complex structures seems to be less useful because it is harder to interpret relationships between broader domains and external criteria. This study provides reasoning for the continued use of current personality instruments that have demonstrated criterion-related validity despite CFA findings that suggest a more complex structure.

The results also refute potential concerns regarding the validity and applicability of previous research based on current personality instruments that has been raised when the inventories failed to be supported by CFA. More importantly, the decrease in construct validity when applying the more complex structure of personality proves that retaining the simple structure of the current questionnaires is not only a defensible option, but may even be favorable.

5. Limitations

In our study, models were specified that reflect the proposed structure according to the respective test manuals and the current typical applications of CFA. Other modeling approaches have been suggested, such as circumplex models (Fabrigar, Visser, & Browne, 1997), and bifactor models that incorporate either method factors (Biderman et al., 2011) or a general factor (Chen, Hayes, Carver, Laurenceau, & Zhang, 2012). These may overcome some issues related to the application of more conventional CFA models to personality data. Their application should be encouraged as they may also provide different views on the internal structure of personality.

The model modifications were based on a data-driven approach. Adding only conceptually sound links between facets and domains may have led to different models that are easier to interpret. It is questionable, however, whether such an arbitrary approach to utilizing CFA results can be defended and is superior to an exclusively conceptual approach to theory development. In any case, it would have led to even fewer additional subscales per broad domain, thus resulting in even smaller relative score changes.

The study did not examine the impact the structures proposed by CFA and ESEM have on criterion-related validity. However, given that the instruments have shown to be less construct-valid, an examination of their criterion-related validity seems not indicated as construct validity should be a requisite before proceeding to this next question.

6. Recommendations and conclusions

Considering the limitations and ambiguities regarding the results obtained from the CFA, one should not dismiss current measures of personality and question their construct validity merely based on the poor fit based on this analytical method. Furthermore, it may be ill-advised to reject personality theory based on CFA results. Theories are designed to explain phenomena and need to simplify the more complex relationships observed between constructs in the real world. Meehl (1990) argues that models can be useful even if they simplify reality. One may add that models *need* to simplify reality so that they can be useful. While current personality measures are not without flaws and do not fulfill the model fit criteria proposed for CFA applications, their continued

use seems justified as they have demonstrated good criterion-related validity. This study also shows that increasing the measures' complexity to comply with CFA standards and improved their internal validity led to a reduced convergent and discriminant validity, suggesting that there is a trade-off between these two aspects of construct validity.

Our results based on ESEM were more promising with regard to the findings on the instruments' construct validity, particularly regarding their discriminant validity. ESEM also offers multi-group analyses and longitudinal analyses, both with tests for measurement invariance (Asparouhov & Muthen, 2009). It hence enables the application of sophisticated methods typically associated with the CFA/structural equation modeling framework but without requiring the instrument to fulfill the more stringent CFA criteria. We believe it to be a useful tool in developing and evaluating self-report questionnaires assessing personality and encourage its application.

Acknowledgments

We would like to thank Lew Goldberg and Maureen Barckley for making the Eugene Springfield Community Sample data available for this study.

References

Aluja, A., Blanch, A., & Garcia, L. F. (2005a). Reanalyzing the 16PF-5 second order structure: Exploratory versus confirmatory factorial analysis. *European Journal of Psychology of Education, 20*(4), 343–353. http://dx.doi.org/10.1007/bf03173561.

Aluja, A., Rossier, J., Garcia, L. F., & Verardi, S. (2005b). The 16PF5 and the NEO-PI-R in Spanish and Swiss samples: A cross-cultural comparison. *Journal of Individual Differences, 26*(2), 53–62. http://dx.doi.org/10.1027/1614-0001.26.2.53.

Asparouhov, T., & Muthen, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling, 16*(3), 397–438. http://dx.doi.org/10.1080/10705510903008204.

Bagozzi, R. P., & Yi, Y. (1991). Multitrait–multimethod matrices in consumer research. *Journal of Consumer Research, 17*(4), 426–439.

Barrick, M. R., & Mount, M. K. (1996). Effects of impression management and self-deception on the predictive validity of personality constructs. *Journal of Applied Psychology, 81*(3), 261–272. http://dx.doi.org/10.1037/0021-9010.81.3.261.

Beauducel, A., & Wittmann, W. W. (2005). Simulation study on fit indexes in CFA based on data with slightly distorted simple structure. *Structural Equation Modeling, 12*(1), 41–75. http://dx.doi.org/10.1207/s15328007sem1201_3.

Biderman, M. D., Nguyen, N. T., Cunningham, C. J. L., & Ghorbani, N. (2011). The ubiquity of common method variance: The case of the Big Five. *Journal of Research in Personality, 45*(5), 417–429. http://dx.doi.org/10.1016/j.jrp.2011.05.001.

Bobko, P., Roth, P. L., & Buster, M. A. (2007). The usefulness of unit weights in creating composite scores. *Organizational Research Methods, 10*(4), 689–709. http://dx.doi.org/10.1177/1094428106294734.

Borkenau, P., & Ostendorf, F. (1990). Comparing exploratory and confirmatory factor analysis: A study on the 5-factor model of personality. *Personality and Individual Differences, 11*(5), 515–524. http://dx.doi.org/10.1016/0191-8869(90)90065-y.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin, 56*(2), 81. http://dx.doi.org/10.1037/h0046016.

Cattell, H. E. P., & Mead, A. D. (2008). The sixteen personality factor questionnaire (16PF®). In G. J. Boyle, G. Matthews, & D. H. Saklofske (Eds.), *The Sage handbook of personality theory and assessment* (pp. 135–159). Los Angeles, CA, USA: SAGE.

Cattell, R. B., & Cattell, H. E. P. (1995). Personality structure and the new fifth edition of the 16PF. *Educational and Psychological Measurement, 55*(6), 926–937. http://dx.doi.org/10.1177/0013164495055006002.

Chen, F. F., Hayes, A., Carver, C. S., Laurenceau, J.-P., & Zhang, Z. (2012). Modeling general and specific variance in multifaceted constructs: A comparison of the bifactor model to other approaches. *Journal of Personality, 80*(1), 219–251. http://dx.doi.org/10.1111/j.1467-6494.2011.00739.x.

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*(1), 93–114. http://dx.doi.org/10.1037/0022-3514.66.1.93.

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

Costa, P. T., & McCrae, R. R. (1992a). 'Four ways five factors are not basic': Reply. *Personality and Individual Differences, 13*(8), 861–865. http://dx.doi.org/10.1016/0191-8869(92)90002-7.

Costa, P. T., & McCrae, R. R. (1992b). *Revised NEO personality inventory (NEO PI-R) and neo five-factor inventory (NEO-FFI)*. FL: Psychological Assessment Resources Odessa.

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology, 41*, 417–440. http://dx.doi.org/10.1146/annurev.ps.41.020190.002221.

Dudley, N. M., Orvis, K. A., Lebiecki, J. E., & Cortina, J. M. (2006). A meta-analytic investigation of conscientiousness in the prediction of job performance: Examining the intercorrelations and the incremental validity of narrow traits. *Journal of Applied Psychology, 91*(1), 40–57. http://dx.doi.org/10.1037/0021-9010.91.1.40.

Fabrigar, L. R., Visser, P. S., & Browne, M. W. (1997). Conceptual and methodological issues in testing the circumplex structure of data in personality and social psychology. *Personality and Social Psychology Review, 1*(3), 184–203. http://dx.doi.org/10.1207/s15327957pspr0103_1.

Fan, X., & Sivo, S. A. (2007). Sensitivity of fit indices to model misspecification and model types. *Multivariate Behavioral Research, 42*(3), 509–529. http://dx.doi.org/10.1080/00273170701382864.

Gerbing, D. W., & Anderson, J. C. (1984). On the meaning of within-factor correlated measurement errors. *Journal of Consumer Research, 11*(1), 572–580.

Goldberg, L. R. (1990). An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology, 59*(6), 1216–1229. http://dx.doi.org/10.1037/0022-3514.59.6.1216.

Goldberg, L. R., Sweeney, D., Merenda, P. F., & Hughes, J. E. (1998). Demographic variables and personality: The effects of gender, age, education, and ethnic/racial status on self-descriptions of personality attributes. *Personality and Individual Differences, 24*(3), 393–403. http://dx.doi.org/10.1016/s0191-8869(97)00110-4.

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*(2), 167–187. http://dx.doi.org/10.1080/00223890701468568.

Hopwood, C. J., & Donnellan, M. B. (2010). How should the internal structure of personality inventories be evaluated? *Personality and Social Psychology Review, 14*(3), 332–346. http://dx.doi.org/10.1177/1088868310361240.

Hu, L.-t., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3*(4), 424–453. http://dx.doi.org/10.1037/1082-989x.3.4.424.

Hu, L.-t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55. http://dx.doi.org/10.1080/10705519909540118.

Hurtz, G. M., & Donovan, J. J. (2000). Personality and job performance: The Big Five revisited. *Journal of Applied Psychology, 85*(6), 869–879. http://dx.doi.org/10.1037/0021-9010.85.6.869.

Judge, T. A., Higgins, C. A., Thoresen, C. J., & Barrick, M. R. (1999). The Big Five personality traits, general mental ability, and career success across the life span. *Personnel Psychology, 52*(3), 621–652. http://dx.doi.org/10.1111/j.1744-6570.1999.tb00174.x.

Kline, P. (1994). *An easy guide to factor analysis*. London: Routledge.

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin, 111*(3), 490–504. http://dx.doi.org/10.1037/0033-2909.111.3.490.

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 indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling, 11*(3), 320–341. http://dx.doi.org/10.1207/s15328007sem1103_2.

Marsh, H. W., Liem, G. A. D., Martin, A. J., Morin, A. J. S., & Nagengast, B. (2011). Methodological measurement fruitfulness of Exploratory Structural Equation Modeling (ESEM): New approaches to key substantive issues in motivation and engagement. *Journal of Psychoeducational Assessment, 29*(4), 322–346. http://dx.doi.org/10.1177/0734282911406657.

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

McCrae, R. R., & John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of Personality, 60*(2), 175–215. http://dx.doi.org/10.1111/j.1467-6494.1992.tb00970.x.

McCrae, R. R., Zonderman, A. B., Costa, P. T., 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*(3), 552–566. http://dx.doi.org/10.1037/0022-3514.70.3.552.

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry, 1*(2), 108–141. http://dx.doi.org/10.1207/s15327965pli0102_1.

Musek, J. (2007). A general factor of personality: Evidence for the Big One in the five-factor model. *Journal of Research in Personality, 41*(6), 1213–1233. http://dx.doi.org/10.1016/j.jrp.2007.02.003.

Muthén, L., & Muthén, B. (2012). *Mplus user's guide version 6.11*. Los Angeles, CA: Muthen & Muthen.

Narayanan, A. (2012). A review of eight software packages for structural equation modeling. *The American Statistician, 66*(2), 129–138. http://dx.doi.org/10.1080/00031305.2012.708641.

Norman, W. T. (1963). Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *The Journal of Abnormal and Social Psychology, 66*(6), 574. http://dx.doi.org/10.1037/h0040291.

Rosseel, Y. (2011). *lavaan: An R package for structural equation modeling and more, version 0.4-11*. <http://lavaan.ugent.be> Accessed January 9, 2012.

Salgado, J. F. (2003). Predicting job performance using FFM and non-FFM personality measures. *Journal of Occupational and Organizational Psychology, 76*(3), 323–346. http://dx.doi.org/10.1348/096317903769647201.

Saris, W. E., Satorra, A., & van der Veld, W. M. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling, 16*(4), 561–582. http://dx.doi.org/10.1080/10705510903203433.

Satorra, A., & Bentler, P. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika, 66*, 507–514. http://dx.doi.org/10.1007/bf02296192.

Schmitt, D. P., Realo, A., Voracek, M., & Allik, J. (2008). Why can't a man be more like a woman? Sex differences in Big Five personality traits across 55 cultures. *Journal of Personality and Social Psychology, 94*(1), 168. http://dx.doi.org/10.1037/0022-3514.94.1.168.

Silver, N. C., & Dunlap, W. P. (1987). Averaging correlation coefficients: Should Fisher's *z* transformation be used? *Journal of Applied Psychology, 72*(1), 146.

Terracciano, A., McCrae, R. R., & Costa, P. T. (2010). Intra-individual change in personality stability and age. *Journal of Research in Personality, 44*(1), 31–37. http://dx.doi.org/10.1016/j.jrp.2009.09.006.

Tett, R. P., Steele, J. R., & Beauregard, R. S. (2003). Broad and narrow measures on both sides of the personality–job performance relationship. *Journal of Organizational Behavior, 24*(3), 335–356. http://dx.doi.org/10.1002/job.191.

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

van der Linden, D., te Nijenhuis, J., & Bakker, A. B. (2010). The general factor of personality: A meta-analysis of Big Five intercorrelations and a criterion-related validity study. *Journal of Research in Personality, 44*(3), 315–327. http://dx.doi.org/10.1016/j.jrp.2010.03.003.

Vassend, O., & Skrondal, A. (2011). The NEO personality inventory revised (NEO-PI-R): Exploring the measurement structure and variants of the five-factor model. *Personality and Individual Differences, 50*(8), 1300–1304. http://dx.doi.org/10.1016/j.paid.2011.03.002.

Ziegler, M., & Buehner, M. (2009). Modeling socially desirable responding and its effects. *Educational and Psychological Measurement, 69*(4), 548–565. http://dx.doi.org/10.1177/0013164408324469.