

## The NEO Personality Inventory Revised: Factor Structure and Gender Invariance From Exploratory Structural Equation Modeling Analyses in a High-Stakes Setting

Adrian Furnham, Nigel Guenole, Steven Z. Levine and Tomas Chamorro-Premuzic

*Assessment* published online 26 July 2012

DOI: 10.1177/1073191112448213

The online version of this article can be found at:

<http://asm.sagepub.com/content/early/2012/07/26/1073191112448213>

---

Published by:



<http://www.sagepublications.com>

**Additional services and information for *Assessment* can be found at:**

**Email Alerts:** <http://asm.sagepub.com/cgi/alerts>

**Subscriptions:** <http://asm.sagepub.com/subscriptions>

**Reprints:** <http://www.sagepub.com/journalsReprints.nav>


**Permissions:** <http://www.sagepub.com/journalsPermissions.nav>

>> [OnlineFirst Version of Record](#) - Jul 31, 2012

[OnlineFirst Version of Record](#) - Jul 26, 2012

[What is This?](#)

# The NEO Personality Inventory–Revised: Factor Structure and Gender Invariance From Exploratory Structural Equation Modeling Analyses in a High-Stakes Setting

Assessment XX(X)  
XX(X) 1–10  
© The Author(s) 2012  
Reprints and permission:  
sagepub.com/journalsPermissions.nav  
DOI: 10.1177/1073191112448213  
http://asm.sagepub.com  


Adrian Furnham<sup>1</sup>, Nigel Guenole<sup>2</sup>,  
Stephen Z. Levine<sup>3</sup>, and Tomas Chamorro-Premuzic<sup>2</sup>

## Abstract

This study presents new analyses of NEO Personality Inventory–Revised (NEO-PI-R) responses collected from a large British sample in a high-stakes setting. The authors show the appropriateness of the five-factor model underpinning these responses in a variety of new ways. Using the recently developed exploratory structural equation modeling (ESEM) technique, the authors show that model fits improve markedly over conventional confirmatory factor analyses (CFA) of the same data set, but that (a) factor interpretations do not change under ESEM analyses, (b) ESEM factor scores, just like CFA factors scores, correlate at near unity with sums of observed scores, (c) NEO-PI-R facets under ESEM analyses are invariant across gender, and (d) ESEM highlights the inappropriateness of alpha and beta as a higher order representation of NEO-PI-R facets, whereas a CFA approach might lead researchers to believe in the appropriateness of these higher order factors. These results, coupled with the existing validity evidence for the NEO-PI-R, suggest that the five-factor structure is the most parsimonious structure for summarizing NEO-PI-R responses from high-stakes settings in the United Kingdom.

## Keywords

NEO Personality Inventory–Revised, NEO-PI-R, Big Five, norms, personality traits, psychometrics

More than half a century since its discovery by early personality researchers, the five-factor model of personality (FFM) is now cemented as the dominant framework for describing consistent differences and similarities between the ways people think, feel, and behave (Chamorro-Premuzic, 2007; Goldberg, 1990, 1993; McCrae & Costa, 1987). The literature underpinning the FFM is overwhelming and the breadth of consensus on the value of the FFM is considerable (Chamorro-Premuzic & Furnham, 2010). This latter point is clearly illustrated by opening paragraphs in numerous journal articles, which contain strikingly similar descriptions of how the FFM model is observed across cultures and languages as well as demographic variables such as gender and age (e.g., Saucier & Goldberg, 1998; Saucier & Ostendorf, 1999). In fact, apart from a few notable exceptions (e.g., Block, 1995, 2001), disputes over the FFM's legitimacy today are centered on issues of refinement rather than questioning the fundamental utility of the model. Some of the aspects that are still being debated relate to the existence or otherwise of a higher order structure of the FFM (e.g., Digman, 1997; Rushton & Irwing, 2009), the status of personality dimensions that are not well represented by the

FFM such as honesty-humility (Ashton & Lee, 2008) and ambition (Hogan & Chamorro-Premuzic, in press), the facet-level structure of the FFM (Perugini & Gallucci, 1997), and whether or not a neuropsychological basis exists for what is essentially a taxonomy of phenotypes (De Young, 2010).

The NEO Personality Inventory–Revised (NEO-PI-R), arguably the leading psychometric measure of the FFM, has played a towering role in bringing researchers to this point in our understanding of personality. Although the research base for the NEO-PI-R is vast, and findings have been consistently replicated in numerous and diverse contexts, the NEO-PI-R model has an Achilles' heel: namely, that confirmatory factor analyses (CFAs) do not yield evidence in

<sup>1</sup>University College London, London, UK

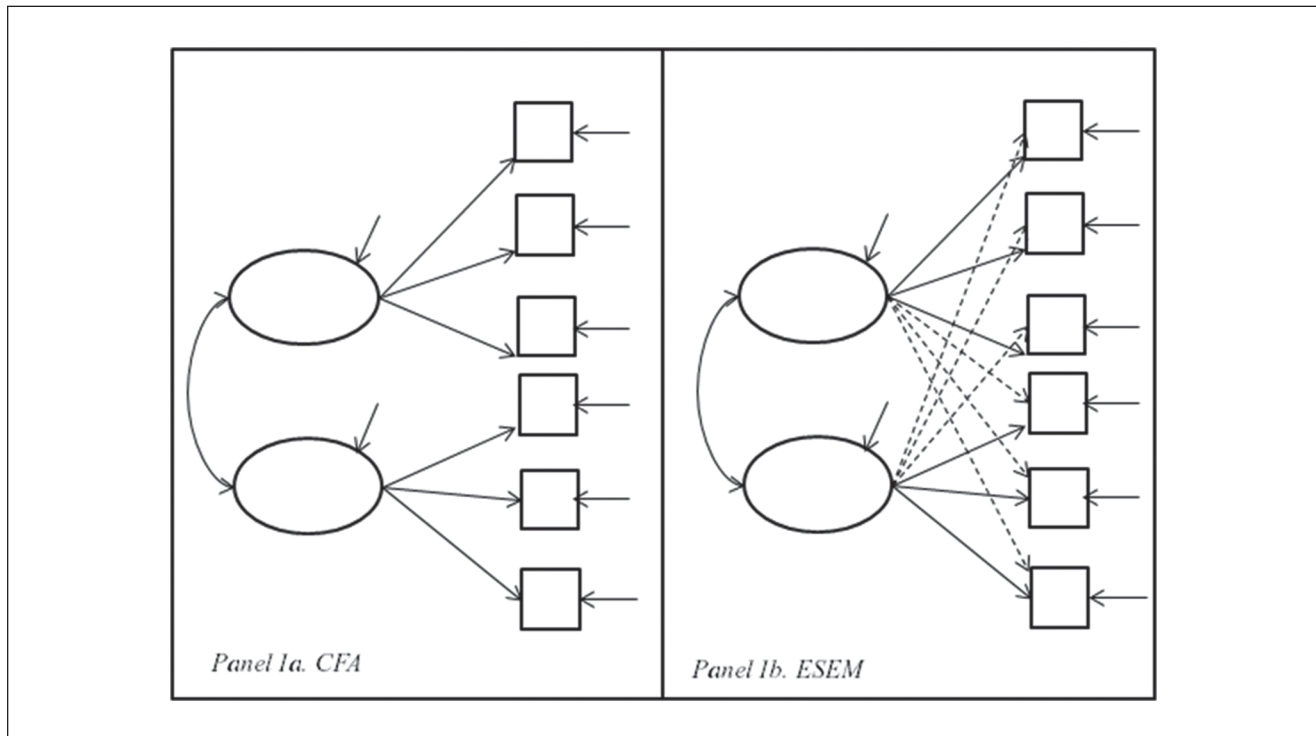
<sup>2</sup>Goldsmiths, University of London, London, UK

<sup>3</sup>Bar Ilan University, Ramat Gan, Israel

## Corresponding Author:

Adrian Furnham, University College London, 26 Bedford Way, London, WC1H 0AP, UK

Email: a.furnham@ucl.ac.uk



**Figure 1.** Confirmatory factor analysis (CFA) versus exploratory structural equation modeling (ESEM) representations of a two-factor model

support of the FFM when judged by traditionally accepted psychometric standards (Church & Burke, 1994; McCrae, Zonderman, Costa, Bond, & Paunonen, 1996). In fact, the fit of CFA models to NEO-PI-R data has routinely been so poor that McCrae et al. (1996) suggested abandoning CFA. The inability of researchers to fit an appropriate CFA model means that there still remain researchers who argue for both fewer (Rushton & Irwing, 2009) and more (Ashton & Lee, 2008) factors of personality. This is in part because of the recognition that inappropriate use of CFA may result in overestimation or underestimation of the number of personality factors extracted.

Exploratory structural equation modeling (ESEM), a recent development in psychological measurement, offers the potential to reconcile these seemingly opposing vantage points. This is because ESEM has a number of advantages over traditional CFA approaches that mean it could be more appropriate for modeling personality data. These advantages include relaxation of the assumption that items have factorial complexity of one (i.e., no cross-loadings of items or facets), the availability of standard errors for parameter estimates in an exploratory setting, and an assessment of fit using goodness-of-fit indices available in traditional structural equation modeling frameworks (Asparouhov & Muthén, 2009; Marsh et al., 2010). The flexibility of ESEM, as contrasted with CFA, is illustrated in Figure 1 for a hypothetical two-factor model. The CFA model in panel 1A

assumes zero loadings on the nontarget factor. On the other hand, the ESEM model in panel 1B allows nonzero loadings on factors other than the primary targeted factor, as illustrated by the dotted arrows. It is thought that this flexibility is more realistic, and these relaxed assumptions are expected to lead to improved fit.

The need for a methodological innovation such as ESEM to reconcile the conflicting evidence from CFA analyses of NEO-PI-R data sets and other methods of validation is well illustrated by research carried out on the short form of the NEO-PI-R. The shorter NEO-FFI (NEO Five-Factor Inventory; 60 vs. 240 items) attracts considerable research interest, no doubt because of its quicker administration time (less than 10 minutes vs. more than 35 minutes for the longer NEO-PI-R). Egan, Deary, and Austin (2000), however, in addition to providing British norms based on 1,025 adults, suggested that the instrument “requires modification and improvement before it can be regarded as measuring the five independent personality traits” (p. 907). In a recent article, Marsh et al. (2010) used the ESEM technique, coupled with theoretically appropriate correlated residuals, to examine fit for the NEO-FFI. Results showed that using ESEM on the NEO-FFI provided a much better model fit than that CFA has exhibited up until this point.

Despite the fact that ESEM has been shown to result in improved model fit for the FFM over traditional CFA approaches, several questions of fundamental importance

to assessment specialists using the NEO-PI-R remain unanswered. First, no studies exist that apply ESEM to longer forms of the NEO-PI-R. As a consequence, there is no evidence as to whether the improved fit observed for the NEO-FFI also exists for the NEO-PI-R. There is also no evidence on whether the FFM primary factor interpretations change as a result of the application of ESEM. Critically, a further issue that remains to be investigated is whether ESEM produces person scores that are well approximated by the simple sums of candidates' observed scores. Because these simple sums are the most routinely interpreted scores in applied settings, the answer to this question has important implications for the accuracy of norm data that is so integral to attributing meaning to personality profiles. Although Marsh, Liem, Martin, Morin, and Nagengast (2011) examined gender invariance for the NEO-FFI, no study has yet examined whether factor scores from ESEM models, with their improved fits, yield FFM scores on the long form NEO-PI-R that are invariant across gender. Finally, we expect researchers and practitioners alike would like to know if an alternative higher order structure exists for more parsimoniously describing personality based on FFM dimensions derived using the more flexible ESEM approach. Several higher order structures of the FFM have been proposed. In a seminal article, Digman (1997) proposed that covariation between the higher order factors of Extraversion and Openness could be explained by a factor he labeled alpha, while covariation among the remaining facets of Neuroticism, Agreeableness, and Conscientiousness could be explained by a factor he called beta. DeYoung (2010) labeled similar higher order structures plasticity and stability based on psychometric as well as neuropsychological evidence. More recently, proponents of a general factor of personality have presented psychometric evidence for what must be the most basic representation of personality so far (Rushton & Irwing, 2009). Given the increased flexibility of ESEM over traditional CFA methods, it would be of interest to practitioners to know whether there is a more parsimonious and well-fitting representation of the FFM that they might use in their applied work based on the factor scores emerging from ESEM. In this article, we answer these questions using a large sample of white-collar British workers who completed the NEO-PI-R in a high-stakes selection setting.

## Method

### Participants

In all, 13,234 British adults were tested over a 10-year period as part of an assessment center run by chartered organizational psychologists. Each participant completed a number of self-report and ability tests as well as other exercises and an interview. Tests were completed for the

purpose of personnel selection or career progression (internal promotion), representing a high-stakes testing scenario as results were used to inform these consequential decisions. Participants' age ranged from 18 to 67 years and most participants were employed as middle or senior managers in a range of British-based companies. Of the 13,234 adults who participated in the assessment center, 4,937 participants completed the NEO-PI-R and were included in the current study, out of whom 25% were female. The average age of the sample was 44 years and the standard deviation was 14 years.

### Measure

The NEO-PI-R (Costa & McCrae, 1992) questionnaire is a 240-item questionnaire designed to measure the FFM traits as well as six primary facets for every trait. Items are responded on a 5-point Likert-type scale that ranged from *strongly disagree* to *strongly agree*. The test is untimed but takes approximately 35 minutes to complete. Although a wealth of research exists providing evidence for the validity and the reliability of this instrument, most data are derived from student and low-stakes settings (Chamorro-Premuzic & Furnham, 2010).

### Procedure

Participants were tested in an assessment center setting for selection and promotion purposes. The questionnaire was untimed and most participants took between 30 and 45 minutes to complete it. They were asked to respond honestly and were promised, and received, full feedback on their scores at a later point.

### Analyses

*Descriptive analyses.* SPSS 17.0 was used for data cleaning. The calculation of the correlations and the simple descriptive analyses are presented in tables later in the text.

*Structural equation models.* All structural equation modeling was carried out using the Mplus computer program (Version 6.1; Muthén & Muthén, 2006) with the MLR estimator to handle issues related to nonnormality of the data. We first fitted both CFA and ESEM models to males and females separately, prior to examining measurement equivalence, which must be demonstrated before concluding that survey items measure constructs similarly across populations. Geomin rotation, an oblique rotation method allowing the factors to be intercorrelated, was performed in the case of ESEM models. The consensus today among methodologists is that oblique rotation should always be preferred. This is because we most constructs in social sciences turn out to be intercorrelated, and an oblique rotation will uncover an orthogonal (i.e., uncorrelated)

solution if one is appropriate anyway (MacCallum, 1998). CFA does not use rotations because the patterns of fixed and free loadings are specified a priori.

**Measurement equivalence.** Measurement equivalence prevails if two individuals with equal standing on the construct assessed, but sampled from different populations (i.e., genders), have equal expected observed scores on the measurement instrument (Drasgow, 1984). Today, numerous publications exist that outline the same steps for showing measurement invariance in multiple group models (e.g., Horn & McArdle, 1992; Millsap, 1995; Vandenberg & Lance, 2000). First, an unconstrained baseline model is estimated where the only parameters equated across genders are those required for identification purposes. If a satisfactory fit for the baseline model is observed, configural, or weak invariance, is said to hold. In other words, the same number of factors exists in the data from both groups and items have the same pattern of zero and nonzero loadings in both groups. Next, factor loadings are constrained to be equal across both groups. If constraints on the factor loadings do not reduce fit appreciably, metric, or strong invariance is said to hold. Next, the intercepts for the items are held equal across groups. If the item intercepts constraints do not appreciably reduce fit, strict, or scalar invariance is said to hold. Although further constraints related to the factor variances, factor covariances, and residuals are possible (cf. Marsh et al., 2011), fulfillment of these additional constraints are not requirements necessary for comparisons of mean differences on the construct under study.

**Model selection.** We report multiple indices in addition to the model chi-square, because its sensitivity to sample size can lead to rejection of theoretically appropriate models (e.g., Byrne, 1998). We selected the most appropriate model out of these sequences of models based on an overall assessment of the following indices: root mean square error of approximation (RMSEA; Steiger, 1990), Bentler's (1990) comparative fit index (CFI), and Bentler and Bonnet's (1980) nonnormed fit index (NNFI). The RMSEA is a measure of badness of fit per degree of freedom, and conventionally values less than .05 are considered indicative of good fit (Brown & Cudeck, 1993). The CFI and Tucker-Lewis index (TLI) range between 0 and 1 and values more than .9 are considered acceptable fit, whereas values more than .95 indicate excellent fit (Hu & Bentler, 1995). Importantly, Marsh, Hau, and Wen (2004) have suggested that these standards are unlikely to be achieved with CFA when models as complex as the FFM are analyzed.

## Results

### Confirmatory Factor Analysis Results

Table 1 presents the results of CFA analyses of facet-level data using Mplus. These results show that the separate

**Table 1.** Fit Statistics for Traditional Confirmatory Factor Analysis Models

CFA	$\chi^2$	df	p	RMSEA	CFI	TLI	SRMR
Male	17640.528	395	<.01	.11	.65	.60	.12
Female	5920.116	395	<.01	.11	.65	.61	.12
Baseline	24293.692	815	<.01	.11	.63	.60	.12

Note. df = degrees of freedom; RMSEA = root mean square error of approximation; CFI = comparative fit index; TLI = Tucker-Lewis index, SRMR = standardized root mean square residual.

**Table 2.** Fit Statistics for Exploratory Structural Equation Models

Model	$\chi^2$	df	p	RMSEA	CFI	TLI	SRMR
Male	4760.87	295	<.01	.06	.91	.86	.03
Female	1634.01	295	<.01	.06	.91	.87	.03
Baseline	6394.87	590	<.01	.06	.91	.86	.03
Metric	6711.34	715	<.01	.06	.90	.88	.04
Scalar	7205.04	740	<.01	.06	.90	.88	.04

Note. df = degrees of freedom; RMSEA = root mean square of approximation; CFI = comparative fit index; TLI = Tucker-Lewis index, SRMR = standardized root mean square residual.

models for males and females fitted the data poorly when compared with the conventional standards outlined above. This pattern of poor fit was also observed for the baseline model for configural invariance. In particular, in addition to chi-square statistics, which are highly significant in all cases, RMSEA is more than .10, which indicates poor model fit. Moreover, CFI and TLI are clearly not anywhere near .90 and .95. Because of this, we did not proceed further to investigate metric and scalar invariance, as CFA models are inappropriate for these data.

### Exploratory Structural Equation Modeling Results

Results from ESEM analyses indicated an improvement in fit for the male-only and female-only models to what can be argued to be acceptable levels by conventionally accepted standards. The improvement in fit between the CFA and ESEM models is approximately comparable in magnitude with the changes that Marsh et al. (2010) observed by moving from CFA to ESEM on the short form of the NEO-PI-R, the NEO-FFI. The better model fit observed here is most likely because of the relaxation of CFA conditions where each facet is only allowed to load on its target factor and has zero loadings on every other factor. Marsh et al. (2010) referred to this CFA model as the independent clusters CFA model. Table 2 shows that for the separate male and female ESEM models, while chi-square remains highly significant; RMSEA and CFI



now indicate adequate fit, whereas TLI is very close to the acceptable level.

### *Measurement Equivalence Results*

The model fit for the baseline model testing configural invariance with ESEM, presented in Table 2, indicates fit statistics similar to the single-group ESEM results. Because this is a substantial improvement over CFA results, and fit statistics approach accepted standards, we proceeded to examine metric and scalar invariance. The results in Table 2 show that the imposition of metric and scalar invariance constraints does not reduce fit appreciably. These findings show that the ESEM solution is a more appropriate solution than the CFA solution for these data, and importantly, that there is no differential facet functioning in this data set across gender. The correlations between the factors for the scalar invariance ESEM solution were small to moderate. The largest correlation, at  $-.45$ , was between Factor I and Factor V. The smallest correlation, at  $-.06$ , was between Factor IV and Factor V.

### *Interpretation of Exploratory Structural Equation Modeling Factors*

One of the primary advantages of the ESEM approach is that it relaxes the assumption that items or facets have zero loadings on all factors other than the target factor. This model restriction is unrealistic for personality data, which are known to be factorially complex (Marsh et al., 2007). Along with the increased flexibility, however, comes the need to interpret the factors as one would in an exploratory factor analyses. In other words, it is quite possible under ESEM that the pattern of factor loadings will not support the a priori FFM structure and patterns of factor loadings need to be examined. Interpretation of the factors that emerge from ESEM ultimately requires judicious interpretation of the loading pattern and significance of the loadings for each of the facets. It seems reasonable, however, to have as our requirement that for a factor to be considered an a priori component of the FFM, all, or at least the majority of the facets that measure the factor ought to have their highest loadings on it, and that all these loadings should be significant.

By these criteria, we see that Factor I is Neuroticism. For both males and females, all the loadings of the N facets are significant on Factor I, and moreover, all facets except N5 Impulsiveness have their highest loadings on Factor I (Table 3). For males, N5 has a higher loading on Factor II and Factor V than it does on Factor I, and only a marginally smaller loading on Factor IV. For females, N5 has a greater loading on Factor V and substantial loadings on Factor II and Factor IV. Constraining these loadings to zero under the independent clusters model that underpins CFA will certainly

detract from model fit. In addition to N5 not having its highest loading on Factor I, which is ostensibly Neuroticism, the facet N6 vulnerability has nontrivial loadings on Factor V for men and women. Using the same criteria, Factor II is Extraversion. All Extraversion facets except E3 assertiveness have their strongest loading on Factor II, and all loadings for Extraversion facets on Factor II are significant for men and women. E3 not only has a greater loading in both the male and female samples on Factor IV than it does on Factor II but it also has a smaller and still considerable loading on Factor V.

The preponderance of evidence points compellingly toward Factor III representing the FFM Openness dimension. For males and females, all Openness facets have their strongest loading on Factor III except O3 feelings, which loads higher on Factor II (that we have labeled Extraversion) in both men and women. Moreover, the need for flexibility in modeling nontarget loadings is again illustrated by sizable secondary loadings for O6 values. Factor IV in the ESEM solution is Agreeableness, because the highest loading for each agreeableness facet is on Factor IV, and all these loadings are significant. Here again, sizable secondary loadings show the inappropriateness of the independent clusters model for these data. Factor V emerges strongly as Conscientiousness with all conscientiousness facets having their highest loadings on the final factor, Factor V, all of which are significant. In sum, the geomin rotated ESEM solutions for men and women reveal two important points. First, ESEM reveals clear support for the FFM. Second, substantial cross-loading for certain facets, most notably N5, E3, and O3, shows the inappropriateness of the independent clusters model assumptions of traditional CFA approaches if the goal is to obtain strong fit from structural equation modeling.

### *Is There a Higher Order Structure Based on the NEO-PI-R's ESEM Solution?*

Although second-order ESEM factor models are not currently possible in Mplus, it is possible to save the factor scores from ESEM and subject these to further analysis using ESEM. We used this procedure to investigate one- and two-factor representations of the factor scores that emerged from the best fitting CFA and ESEM models described earlier. First, the single-factor model based on CFA scores did not converge, suggesting that a single-factor model is inappropriate for these data. The single-factor model for the ESEM scores did converge, but fit was so bad as to preclude further interpretation (RMSEA = .32, CFI = .71, TLI = .43). The two-factor model based on CFA factor scores did converge. The factors clearly resembled alpha or plasticity with high Extraversion and Openness loadings ( $N = -.01$ ,  $E = .77$ ,  $O = .87$ ,  $A = .21$ ), and beta or stability with high-reverse Neuroticism, Agreeableness, and

**Table 3.** Loading Parameter Estimates and Significance From Standardized ESEM Solution

Facet	Male										Female									
	Est.	p	Est.	p	Est.	p	Est.	p	Est.	p	Est.	p	Est.	p	Est.	p	Est.	p		
N1	.79	.00	.01	.54	.08	.00	.03	.05	.00	.96	.82	.00	.01	.54	.07	.00	.03	.05	.00	.96
N2	.65	.00	.02	.39	.01	.47	-.48	.00	-.03	.24	.71	.00	.01	.40	.01	.47	-.46	.00	-.03	.24
N3	.73	.00	-.05	.06	.10	.00	.00	.75	-.18	.00	.74	.00	-.04	.07	.09	.00	.00	.75	-.16	.00
N4	.61	.00	-.17	.00	.01	.53	.09	.00	-.06	.03	.62	.00	-.14	.00	.01	.53	.08	.00	-.06	.03
N5	.34	.01	.40	.00	-.01	.49	-.25	.00	-.41	.00	.37	.01	.36	.00	.00	.49	-.24	.00	-.39	.00
N6	.54	.00	-.03	.39	-.05	.01	.11	.00	-.36	.00	.56	.00	-.02	.39	-.05	.01	.10	.00	-.33	.00
E1	.00	.94	.84	.00	.00	.76	.35	.00	.09	.06	.00	.94	.84	.00	.00	.76	.36	.00	.09	.06
E2	-.05	.48	.68	.00	-.13	.00	.06	.03	.02	.62	-.05	.48	.66	.00	-.13	.00	.06	.03	.02	.62
E3	-.15	.00	.32	.00	.00	.68	-.42	.00	.27	.00	-.17	.00	.29	.00	.00	.68	-.40	.00	.26	.00
E4	-.01	.78	.33	.00	.09	.00	-.31	.00	.32	.00	-.01	.78	.31	.00	.09	.00	-.30	.00	.32	.00
E5	-.05	.40	.38	.00	.10	.00	-.20	.00	-.10	.00	-.05	.40	.33	.00	.10	.00	-.18	.00	-.10	.00
E6	-.04	.34	.69	.00	.13	.00	.09	.00	-.02	.59	-.05	.34	.70	.00	.14	.00	.09	.00	-.02	.59
O1	.01	.79	.19	.00	.52	.00	-.06	.00	-.37	.00	.01	.79	.18	.00	.53	.00	-.06	.00	-.37	.00
O2	-.01	.87	.00	.48	.67	.00	.13	.00	-.06	.01	-.01	.87	.00	.49	.71	.00	.13	.00	-.06	.01
O3	.31	.00	.51	.00	.39	.00	-.01	.56	.03	.43	.36	.00	.51	.00	.41	.00	-.01	.55	.03	.43
O4	-.26	.00	.24	.00	.39	.00	.00	.91	-.12	.00	-.30	.00	.23	.00	.41	.00	.00	.91	-.13	.00
O5	-.21	.02	-.14	.00	.70	.00	.01	.49	.01	.39	-.23	.02	-.13	.00	.70	.00	.01	.50	.01	.40
O6	-.20	.00	.16	.00	.27	.00	.01	.47	-.16	.00	-.24	.00	.17	.00	.29	.00	.01	.48	-.17	.00
A1	-.24	.00	.34	.00	.03	.11	.42	.00	-.02	.24	-.26	.00	.31	.00	.03	.11	.40	.00	-.02	.24
A2	-.01	.87	.04	.13	-.05	.00	.55	.00	.11	.00	-.01	.87	.04	.13	-.05	.00	.53	.00	.11	.00
A3	.06	.29	.55	.00	.00	.81	.62	.00	.24	.00	.07	.29	.52	.00	.00	.81	.61	.00	.25	.00
A4	-.20	.00	.03	.42	.02	.30	.73	.00	-.01	.53	-.23	.00	.03	.42	.02	.30	.72	.00	-.01	.54
A5	.09	.01	-.01	.81	-.03	.10	.47	.00	-.04	.09	.10	.01	-.01	.81	-.03	.10	.47	.00	-.04	.09
A6	.09	.03	.22	.00	.14	.00	.53	.00	.01	.47	.11	.03	.21	.00	.15	.00	.53	.00	.01	.47
C1	-.16	.23	.03	.33	.14	.00	-.02	.26	.64	.00	-.18	.23	.02	.34	.13	.00	-.01	.26	.62	.00
C2	.21	.11	-.08	.00	-.02	.29	-.01	.41	.67	.00	.23	.11	-.07	.00	-.02	.29	-.01	.41	.65	.00
C3	.04	.79	-.01	.57	-.05	.01	.14	.00	.76	.00	.05	.79	-.01	.57	-.05	.01	.14	.00	.74	.00
C4	.04	.75	.07	.05	.12	.00	-.24	.00	.68	.00	.04	.75	.07	.05	.12	.00	-.23	.00	.67	.00
C5	-.09	.55	-.01	.62	.02	.24	.00	.99	.77	.00	-.10	.55	-.01	.62	.02	.24	.00	.99	.73	.00
C6	-.05	.78	-.35	.00	.00	.80	.24	.00	.62	.00	-.05	.78	-.31	.00	.00	.80	.22	.00	.58	.00

Note. Est. = estimated geomin rotated factor loading; p = two-tailed p value. N = Neuroticism; N1 = Anxiety; N2 = Angry Hostility; N3 = Depression; N4 = Self-Consciousness; N5 = Impulsiveness; N6 = Vulnerability; E = Extraversion; E1 = Warmth; E2 = Gregariousness; E3 = Assertiveness; E4 = Activity; E5 = Excitement Seeking; E6 = Positive Emotion; O = Openness to Experience; O1 = Fantasy; O2 = Aesthetics; O3 = Feelings; O4 = Actions; O5 = Ideas; O6 = Values; A = Agreeableness; A1 = Trust; A2 = Straightforwardness; A3 = Altruism; A4 = Compliance; A5 = Modesty; A6 = Tender Mindedness; C = Conscientiousness; C1 = Competence; C2 = Order; C3 = Dutifulness; C4 = Achievement Striving; C5 = Self-Discipline; C6 = Deliberation.

Conscientiousness loadings (N = -.83, E = .45, O = -.01, A = .18, and C = .80). Although we might conclude the existence of higher order factors on the basis and fit (RMSEA = .13, CFI = .99, TLI = .91), something is clearly wrong at a chi-square of 88.86 and a single degree of freedom. Moreover, these values inspected casually would disguise the fact that the values are based on an even worse fitting first-order model (i.e., the multiple-group CFA baseline model presented in Table 1). Despite evidence of alpha and beta based on loading patterns, the model fit and knowledge that the factor solution is based on a poorly fitting first-order solution demands caution before interpreting the supposed alpha and beta factors as substantive factors. Importantly, the two-factor ESEM solution that analyzed

the first-order ESEM factors would not converge because of nonpositive definite covariance matrix, suggesting that the higher order solution was not appropriate. Thus, the application of ESEM highlights the inappropriateness of a higher order solution for the Big Five based on these data. The rationale we offer for this result is an intriguing explanation suggested by Ashton, Lee, Goldberg, and De Vries (2009). The essence of the argument of Ashton et al. is that suppressed secondary loadings on nontarget constructs can lead to correlations between factors. Although these correlations among factors can be modeled by higher order factors, the higher order factors accounting for these correlations will be spurious, because the correlations on which they are based are artifactual. On the basis of these

**Table 4.** Correlations Among ESEM, CFA, and Observed Variable Scores

Domain	CFA–OBS	ESEM–OBS	CFA–ESEM
Neuroticism	.97	.96	.97
Extraversion	.94	.93	.95
Openness	.96	.97	.92
Agreeableness	.96	.84	.80
Conscientiousness	.96	.97	.98

Note. CFA–OBS = Correlation between confirmatory factor analysis factor scores and corresponding observed variable sums; ESEM–OBS = correlation between exploratory structural equation modeling factor scores and corresponding observed variable sums; CFA–ESEM = correlations between confirmatory factor analysis factor scores and exploratory structural equation modeling factor scores.

results, and if well-fitting models are of concern to practitioners, the most general level of the FFM that yields adequate and defensible fit, so far, is at the level of the five factors.

### Correlations Between ESEM Factors, CFA Factors, and Observed Variable Totals

The way that norms for the NEO-PI-R are typically used, as is the case with other personality tests, involves creating observed item totals for candidates and calculating the proportion of a representative norm sample that scores the same as or lower than the candidate. It is today well known that correlations between latent variable scores and observed variable totals are often very high (Fan, 1998). Correlations between latent variable scores and external criteria and observed variable scores and external criteria are also known to be similar (Ferrando & Chico, 2007), justifying use of observed variable scores in the scoring and feedback process. However, assessment practitioners who use the NEO-PI-R will want to know whether the observed variable scores they routinely use are still appropriate, now that a more flexible latent variable modeling approach has yielded better model fit to the NEO-PI-R.

To investigate this issue, we saved the factor scores from both the poor fitting CFA models along with the better fitting ESEM models, and correlated each set of scores both with each other and with their observed variable equivalents. Results, presented in Table 4, indicated that the correlations between both forms of latent scores (i.e., ESEM factor scores or CFA factor scores) and observed variable scores were near unity in almost all instances. This suggests that practitioners may continue to use observed scores in their work because the improved model fit for the ESEM scores does not substantially affect the scores that emerge from these analyses. The exception to this pattern is for the

Agreeableness factor, where the correlations are still in excess of .90 for the CFA-observed correlations, but the correlations between the ESEM and CFA and ESEM and observed variable scores drop to .80 and .84, respectively. However, we expect that even these correlations of .80 and above will give many practitioners the confidence to continue using observed variable scores for even the Agreeableness dimension of the NEO-PI-R.

### Demographic Descriptive for the Current British Sample

Applied measurement specialists reading this article might be interested in descriptive statistics for this sample to enable them to calculate norms for interpretation in their own work. They might also be interested to examine gender differences or age-related patterns of association with personality. In response, in Table 5, we present means for the overall sample as well as for males and females separately. Effect sizes, calculated as Cohen's *d*, are presented for gender. These show that overall the effect sizes are small to moderate, with the largest effect size at the factor level occurring for the Openness domain ( $d = .52$  favoring females) whereas the largest facet-level effect size occurred on O3 feelings ( $d = .55$  favoring females). The magnitude of the differences observed, where just a handful of facets show moderate-sized differences of a half standard deviation, mirrors the findings of Costa, Terracciano, and McCrae (2001). Costa et al. reported that the largest gender difference for the NEO-PI-R was .44 for the N6 vulnerability facet of Neuroticism. Other similarities also exist between our study and theirs, for example, females are uniformly higher on all Neuroticism facets in both studies. The United Kingdom, however, was not represented in Costa et al.'s study. We also recommend caution in making comparisons as it is today clear that the purpose of the personality testing has an impact on the mean levels observed for personality measures (e.g., De Fruyt et al., 2006). That is to say, score inflation in selection settings is likely to render comparisons with data sets from lower stakes personality testing situations inappropriate.

Table 5 contains the correlation between the NEO-PI-R domain scores, along with its facets, and age. These results show that, by and large, there are only weak associations between personality and age. Where significant correlations between age and personality are observed, for example with agreeableness, the results show that the correlations are in fact quite small. This means that there is only a very minor association between personality and age. Moreover, significant associations when the correlations are so small suggest that they are the result of high power from the large sample size, rather than showing a substantively meaningful relationship between age and personality.



**Table 5.** Sample Descriptive Statistics for NEO-PI-R Facets

Scale	Overall		Male (n = 3,715)		Female (n = 1,222)		Cohen's d Effect Size			Age		
	r	Mean	SD	Mean	SD	Mean	SD	d	95% LB	95% UB	r	p
N		65.56	19.42	63.98	18.99	70.39	19.95	-.33	-.87	.20	.01	.79
N1		12.44	5.14	12.01	5.02	13.75	5.28	-.34	-.48	-.20	.01	.51
N2		10.17	4.53	10.02	4.49	10.63	4.60	-.13	-.26	-.01	.01	.57
N3		9.69	4.66	9.44	4.55	10.42	4.92	-.21	-.34	-.10	-.02	.36
N4		11.86	4.22	11.68	4.15	12.42	4.38	-.18	-.29	-.06	.00	.90
N5		14.75	4.36	14.44	4.34	15.68	4.30	-.28	-.41	-.16	.00	.86
N6		6.72	3.46	6.42	3.38	7.62	3.54	-.35	-.44	-.25	.02	.47
E		127.14	18.22	126.12	18.37	130.23	17.42	-.23	-.73	-.28	-.04	.06
E1		23.73	3.93	23.22	3.94	25.30	3.48	-.54	-.65	-.44	-.01	.62
E2		20.05	4.63	19.79	4.67	20.84	4.41	-.23	-.36	-.10	-.01	.52
E3		20.55	4.55	20.73	4.49	19.98	4.66	.17	.04	.29	-.01	.64
E4		21.54	4.12	21.44	4.15	21.85	3.99	-.10	-.22	-.01	-.02	.43
E5		18.77	4.49	18.89	4.44	18.42	4.63	-.11	-.02	.23	-.06	.00
E6		22.51	4.58	22.07	4.62	23.82	4.19	-.39	-.51	-.26	-.04	.07
O		120.96	18.52	118.63	18.24	128.05	17.54	-.52	-1.03	-.02	-.02	.49
O1		16.93	4.78	16.55	4.76	18.06	4.65	-.32	-.45	-.19	.00	.84
O2		17.76	5.95	17.11	5.96	19.75	5.46	-.45	-.62	-.29	.01	.58
O3		21.84	4.19	21.28	4.17	23.52	3.81	-.55	-.66	-.44	-.02	.37
O4		20.10	4.18	19.67	4.20	21.44	3.83	-.43	-.55	-.32	-.01	.77
O5		20.54	5.25	20.39	5.26	21.00	5.20	-.12	-.26	.03	-.01	.54
O6		23.79	3.42	23.65	3.48	24.22	3.17	-.17	-.26	-.07	-.03	.18
A		118.53	15.67	117.13	15.65	122.79	14.96	-.37	-.80	.07	-.05	.01
A1		21.74	4.06	21.63	4.04	22.07	4.11	-.11	-.22	.01	-.04	.09
A2		18.36	4.45	18.10	4.46	19.16	4.31	-.24	-.36	-.12	-.06	.01
A3		23.68	3.55	23.30	3.56	24.83	3.27	-.44	-.53	-.34	-.03	.11
A4		18.15	4.01	17.99	4.03	18.64	3.93	-.16	-.27	-.05	-.01	.65
A5		17.10	4.55	16.85	4.58	17.88	4.37	-.23	-.36	-.10	-.01	.58
A6		19.49	3.50	19.27	3.54	20.18	3.29	-.26	-.36	-.16	-.05	.03
C		132.97	17.38	133.32	17.41	131.91	17.24	-.08	-.40	-.57	.00	.84
C1		24.05	3.23	24.15	3.19	23.76	3.30	.12	.03	.21	.02	.40
C2		18.97	4.51	18.87	4.52	19.27	4.47	-.09	-.21	-.04	-.01	.63
C3		24.65	3.56	24.77	3.54	24.28	3.60	.14	.04	.24	-.02	.35
C4		22.97	4.02	23.01	4.04	22.84	3.98	.04	-.07	.15	-.03	.23
C5		23.81	4.05	23.78	4.04	23.93	4.09	-.04	-.15	.07	-.02	.35
C6		18.54	4.41	18.77	4.36	17.84	4.50	.21	.09	.33	.03	.12

Note. NEO-PI-R = NEO Personality Inventory-Revised; LB = lower bound; UB = upper bound; N = Neuroticism; N1 = Anxiety; N2 = Angry Hostility; N3 = Depression; N4 = Self-Consciousness; N5 = Impulsiveness; N6 = Vulnerability; E = Extraversion; E1 = Warmth; E2 = Gregariousness; E3 = Assertiveness; E4 = Activity; E5 = Excitement Seeking; E6 = Positive Emotion; O = Openness to Experience; O1 = Fantasy; O2 = Aesthetics; O3 = Feelings; O4 = Actions; O5 = Ideas; O6 = Values; A = Agreeableness; A1 = Trust; A2 = Straightforwardness; A3 = Altruism; A4 = Compliance; A5 = Modesty; A6 = Tender Mindedness; C = Conscientiousness; C1 = Competence; C2 = Order; C3 = Dutifulness; C4 = Achievement Striving; C5 = Self-Discipline; C6 = Deliberation.

## Discussion

Although the NEO-PI-R is one of the most well researched instruments available for the assessment of broad dimensions of personality, and large norm databases exist, researchers

have not been able to show satisfactory model fit for it using modern psychometric methods such as CFA. The current results highlight that CFA, the most common method to analyze the FFM factors, makes unrealistic assumptions with regard to the factorial complexity of each

of the NEO-PI-R facets. These facets clearly have nonzero loadings on numerous factors, and this violates what Marsh et al. (2007) referred to as the independent clusters model. Importantly, we showed that the factors retain their a priori interpretations when modeled at facet level using ESEM. The factor scores estimated based on the well-fitting ESEM model were then shown to be correlated almost perfectly with both the scores estimated based on the CFA model and their observed variable total counterparts. The biggest discrepancy occurred for agreeableness, although correlations were still large (.80 with CFA scores and .84 with ESEM scores). The ESEM model also showed measurement equivalence for the number of factors (configural invariance), facet loadings (metric invariance), and facet intercepts (scalar invariance). This suggests that the relation between facet-level scores and the latent personality dimension is the same for both males and females. Finally, analyses of the ESEM factors revealed a structure that resembled Digman's (1997) alpha and beta for ESEM analyses of factor scores derived from CFA, but this model fitted poorly, and was based on factor scores from a first-order model that had even worse fit. Under ESEM analyses of first-order factor scores derived from a well-fitting first-order model, a two-factor model was shown to be inappropriate because of inadmissible solutions. The most general level of interpretation appropriate for responses to the NEO-PI-R, therefore, is still the FFM if model fit is of concern.

Limitations of this study ought to be mentioned to facilitate interpretation of results and guide future research. First, these data are collected in a setting where individuals were being considered for selection and promotion. Results observed here might not generalize to other situations in which personality might be measured (e.g., research purposes, or solely for development). Care should be taken then when examining the consistency between these findings and other studies using data from other contexts. Our study used ESEM to model facet-level data, and it would be useful to apply the same analyses at the item level for the NEO-PI-R. Finally, although ESEM has not revealed any strong implications for changing applied practice, we have only so far examined measurement models (i.e., modeled responses to the questionnaire itself). It might be that when ESEM is used in a broader setting that included criterion variables we will identify criteria for which the better fit yields predictive improvements. Future research should investigate this issue.

Because of the unique nature and characteristics of this sample we also presented descriptive statistics and examined gender differences and age associations for each of the NEO-PI-R factors and facets. These show that gender differences are small, and age-related associations are weak. Indeed, it is because past research has shown only small demographic associations with personality (Sackett, Schmitt, Ellingson, & Kabin, 2001) and still predicts performance

that, tests such as the NEO-PI-R are used so widely in occupational settings. Because the gender differences were small and associations with age were weak, the descriptive statistics presented here might be used by industrial-organizational psychologists wanting to create norms for high-stakes selection in the United Kingdom without too much concern over impact against women and older workers at mid- to advanced career stages.

### Declaration of Conflicting Interests

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

### Funding

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

### References

- Ashton, M. C., & Lee, K. (2008). The prediction of honesty-humility-related criteria by the HEXACO and five-factor models of personality. *Journal of Research in Personality, 42*, 1216-1228.
- 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.
- Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*, 238-246.
- Bentler, P. M., & Bonnet, D. C. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*, 588-606.
- Block, J. (1995). A contrarian view of the five-factor approach to personality description. *Psychological Bulletin, 117*, 187-215.
- Block, J. (2001). Millennial contrarianism: The five-factor approach to personality description 5 years later. *Journal of Research in Personality, 35*, 98-107.
- Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Beverly Hills, CA: Sage.
- Byrne, B. (1998) *Structural Equation Modeling With Lisrel, Prelis, and Simplis: Basic Concepts, Applications, and Programming* (Multivariate Applications Book Series). Lawrence Erlbaum Associates.
- Chamorro-Premuzic, T. (2007). *Personality and individual differences*. Oxford, England: Blackwell-Wiley.
- Chamorro-Premuzic, T., & Furnham, A. (2010). *The psychology of personnel selection*. Cambridge, England: Cambridge University Press.
- Costa, P. T., & McCrae, R. R. (1992). *The NEO PI-R professional manual*. Odessa, FL: Psychological Assessment Resources.
- Costa, P., Terracciano, A., & McCrae, R. (2001). Gender differences in personality traits across cultures: Robust and surprising findings. *Journal of Personality and Social Psychology, 81*, 322-331.

- 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.
- De Fruyt, F., Bartels, M., Van Leeuwen, K. G., De Clercq, B., Decuyper, M., Mervielde, I. (2006). Five types of personality continuity in childhood and adolescence. *Journal of Personality and Social Psychology*, *913*, 538-552.
- DeYoung, C. (2010). Personality neuroscience and the biology of traits. *Social and Personality Psychology Compass*, *4*, 1165-1180.
- Digman, J. M. (1997). Higher-order factors of the big five. *Journal of Personality and Social Psychology*, *73*, 1246-1256.
- Drasgow, F. (1984). Scrutinizing psychological tests: Measurement equivalence and equivalent relations with external variables are the central issues. *Psychological Bulletin*, *95*, 134-135.
- Egan, V., Deary, I., & Austin, E. (2000). The NEO-FFI: Emerging British norms and an item level analysis suggest N, A and C are more reliable than O and E. *Personality and Individual Differences*, *29*, 907-920.
- Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement*, *58*, 357-381.
- Ferrando, P., & Chico, E. (2007). The external validity of scores based on the two-parameter logistic model: Some comparisons between IRT and CTT. *Psicológica*, *28*, 237-257.
- Goldberg, L. R. (1990). An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology*, *59*, 1216-1229.
- Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist*, *48*, 26-34.
- Hu, L., & Bentler, P. M. (1995). Evaluating Model Fit. In R.H. Hoyle (Ed.), *Structural Equation Modeling Concepts Issues and Applications*, (pp. 76-99). Thousand Oaks, CA: Sage.
- Hogan, R., & Chamorro-Premuzic, T. (in press). Personality and career success. In L. Cooper et al. (Eds.). *Handbook of personality and social psychology*. Washington, DC: APA.
- Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance for aging research. *Experimental Aging Research*, *18*, 117-144.
- MacCallum, R. (1998). Commentary on quantitative methods in I/O research. *Industrial-Organizational Psychologist*, *35*, 4.
- 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 over-generalising Hu & Bentler's (1999) findings. *Structural Equation Modeling*, *11*, 320-341.
- 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*, 322-346.
- Marsh, H. W., Lüdtke, O., Muthén, B., Asparouhov, T., Morin, A. J. S., Trautwein, U., & Nagengast, B. (2010). A new look at the big-five factor structure through exploratory structural equation modeling. *Psychological Assessment*, *22*, 471-491.
- McCrae, R. R., & Costa, P. T., Jr. (1987). Validation of the five-factor model across instruments and observers. *Journal of Personality and Social Psychology*, *52*, 81-90.
- 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.
- Millsap, R. E. (1995). Measurement invariance, predictive invariance, and the duality paradox. *Multivariate Behavioral Research*, *30*, 577-605.
- Muthén, L., & Muthén, B. (2006). *Mplus user's guide*. Los Angeles, CA: Muthén & Muthén.
- Perugini, M., & Gallucci, M. (1997). A hierarchical faceted model of the Big Five. *European Journal of Personality*, *11*, 279-301.
- Rushton, J. P., & Irwing, P. (2009). A General Factor of Personality (GFP) from the Multidimensional Personality Questionnaire. *Personality and Individual Differences*, *47*, 571-576.
- Sackett, P. R., Schmitt, N., Ellingson, J. E., & Kabin, M. B. (2001). High-stakes testing in employment, credentialing, and higher education: Prospects in a post-affirmative action world. *American Psychologist*, *56*, 302-318.
- Saucier, G., & Goldberg, L. R. (1998). What is beyond the Big Five? *Journal of Personality*, *66*, 495-524.
- Saucier, G., & Ostendorf, F. (1999). Hierarchical subcomponents of the Big Five personality factors: A cross-language replication. *Journal of Personality and Social Psychology*, *76*, 613-627.
- Steiger, J. H. (1990). Structural model evaluation and modification. *Multivariate Behavioral Research*, *25*, 173-180.
- Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, 4-70.