Testing the factor structure and measurement invariance across gender of the Big Five Inventory through exploratory structural equation modeling

Carlo Chiorri[1,2], Herbert W. Marsh[3], Alessandro Ubbiali[2], & Deborah Donati[2]

[1] Department of Educational Sciences, Psychology Unit, University of Genova, Italy

[2] Psyche-Dendron Association, Italy

[3] Institute for Positive Psychology and Education, Australian Catholic University, Strathfield, Australia; Department of Education, University of Oxford, Oxford, England; and School of Education, King Saud University, Riyadh, Saudi Arabia

Correspondence to:
Carlo Chiorri, PhD
Department of Educational Sciences
University of Genova
Corso Podestà, 2
16128 Genova (Italy)
Italy.
Tel. + 39 010 209 53709
Fax + 39 010 209 53728
E-mail address: carlo.chiorri@unige.it

PRE-PRINT: This is a pre-print version of the accepted manuscript and as such may contain errors. The paper is under copyright and the final version must be cited as:

Chiorri, C. , Marsh, H. W., Ubbiali, A., & Donati, D. (2016). Testing the factor structure and measurement invariance across gender of the Big Five Inventory through Exploratory Structural Equation Modeling. *Journal of Personality Assessment*, *98*(1), 88-99. doi: 10.1080/00223891.2015.1035381.

Abstract

Confirmatory factor analyses (CFAs) typically fail to support the a priori five-factor structure of Big Five self-report instruments, due in part to the overly restrictive CFA assumptions. We show that exploratory structural equation modeling (ESEM), an integration of CFA and exploratory factor analysis, overcomes these problems in relation to responses to the 44-item Big Five Inventory (BFI) administered to a large Italian community sample. ESEM fitted the data better and resulted in less correlated factors than CFA, although ESEM and CFA factor scores correlated at near unity with observed raw scores. Tests of gender invariance with a 13-model taxonomy of full measurement invariance showed that the factor structure of the BFI is gender-invariant and that women score higher on neuroticism, agreeableness, extraversion and conscientiousness. Through ESEM one could address substantively important issues about BFI psychometric properties that could not be appropriately addressed through traditional approaches.

Keywords: Exploratory Structural Equation Modeling, Five Factor Approach, Big Five Inventory, Confirmatory Factor Analysis, Measurement invariance

Testing the factor structure and measurement invariance across gender of the Big Five Inventory through exploratory structural equation modeling

Introduction

Arguably, a major breakthrough in personality psychology in the last few decades has been the emergence and acceptance of the so-called Five-Factor Approach to personality (FFA; as in Block, 2010, we use this generic term, since it is not related to any specific group of researchers or instruments). This approach assumes that individual differences in adult personality characteristics can be organized in terms of five broad trait domains: extraversion, agreeableness, conscientiousness, neuroticism, and openness to experience. A huge body of empirical research has supported the stability and predictive validity of the FFA factors across different populations, settings and countries (e.g., McCrae & Costa, 1997): this suggests that the FFA factors are a universal language in personality psychology. Nevertheless, contrarian views do exist: these argue that the self-report FFA factors do not necessarily provide an adequate representation of global personality (e.g., Block, 2010).

Several measures of the FFA factors have been developed. John, Donahue, and Kentle (1991) addressed the need for a relatively short instrument measuring the prototypical components of the Big Five by developing the Big Five Inventory (BFI). Their aim was to create a brief inventory that would allow quick, efficient, and flexible assessment of the FFA factors when there is no need for more differentiated measurement of individual facets (John, Naumann, & Soto, 2008). They used 44 short phrases that included trait adjectives known to be prototypical markers of the Big Five, to be rated on a 5-point, Likert-type agreement scale. These adjectives were accompanied by elaborative, clarifying, or contextual information (John et al., 2008). The BFI scales have shown adequate internal consistency, test-retest reliability, and clear factor structure; they have also shown substantial convergence with longer Big Five measures (e.g., Benet-Martinéz & John, 1998; John et al., 2008).

The BFI has been translated into 29 different languages and administered into 56 nations; the five-dimensional structure has proved robust across major regions of the world (Schmitt, Allik, McCrae, & Benet-Martinez, 2007). Adaptation studies showed that the sound psychometric properties of the English original were retained at least in the Spanish (Benet-Martinez & John, 1998), German (Lang, Lüdtke, & Asendorpf, 2001), Dutch (Denissen, Geenen, van Aken, Gosling, & Potter, 2008), French (Plaisant, Courtois, Réveillère, et al., 2010) and Italian (Ubbiali, Chiorri, Hampton, & Donati, 2013) versions. All these studies found that a varimax-rotated principal component analysis (PCA) yielded a 5-component simple structure, with substantial loadings on the target factors and minimal cross-loadings. However, Cid and Finney (2009) and Ubbiali et al. (2013) argued that PCA may not be an appropriate data reduction method, since the BFI items are thought to be operationalizations of latent constructs: as a consequence, the interest is in the common variance among them. Hence, the measurement model to be specified should be a reflective indicator model (Bollen & Lennox, 1991), which would be better tested through exploratory factor analysis (EFA). However, Ubbiali et al. (2013) reported that, in the case of BFI, EFA and PCA yielded overlapping results, apparently supporting the claim of Velicer and Jackson (1990) that the exploratory data reduction method is unlikely to have any substantial effect on empirical results or conclusions.

Despite the large body of empirical research supporting the robustness of the 5-factor structure of the BFI, almost all studies investigating its psychometric properties have relied on exploratory (EFA or PCA), rather than confirmatory, analyses. The main reason for this approach seems to be that confirmatory factor analyses (CFA) have invariably failed to provide clear support for the five-factor model, regardless of the Big Five measures employed (e.g., Vassend & Skrondal, 1997 with the NEO-PI-R; Cooper, Smillie, & Corr, 2010 with the Mini-IPIP).

Similar results also were obtained with the BFI. Benet-Martinez and John (1998) carried out a multiple-group CFA (MG-CFA) to test the measurement invariance of the BFI among English- and Spanish-speaking participants, and specified two models with invariant factor loadings, one with uncorrelated and one with correlated factors. They found that Model 1 did not adequately fit the data, while Model 2 did (using the model chi-square-to-degrees of freedom ratio [2.11] and the Comparative Fit Index [CFI = .92] as measures of goodness of fit). Despite the seemingly sound results, it must be noted that they did not specify the classic independent clusters model (ICM) usually employed in CFA studies, which requires each indicator to load on only one factor and all cross-loadings to be zero, but allowed the estimation of two cross-loadings, which were further constrained to be equal among groups.

Levine and Jackson (2002) performed a CFA on data from 153 English employees and found what they considered a good fit, but reported only the chi-square to degrees of freedom ratio (1.70) and the Root Mean Square Error of Approximation (RMSEA = .06). It should be noted, however, that testing the factor structure of the BFI through CFA was not the primary aim of their study. Nor was this the case with Vandeberghe, St-Onge, and Robineau (2008), who administered the BFI to 967 Quebecer professionals. Though they used the most common fit indices—the CFI, the Non-Normed Fit Index (NNFI), the Tucker-Lewis Index (TLI) and the RMSEA, they assessed item parcels instead of all items together. That is, they combined the items to create three aggregated indicators for each dimension. They considered this strategy to be justified “given the complexity of the model under evaluation” (Vandeberghe et al., 2008, p. 435). Leaving aside the controversy over the actual utility and efficacy of parcels (e.g., Little, Rhemtulla, Gibson, & Schoemann, 2013; Marsh, Lüdtke, Nagengast, Morin, & Von Davier, 2013), the fit of the model was only marginally acceptable (NNFI = .88, CFI = .89, RMSEA = .08). Chiorri, Ubbiali, and Donati (2008) found that a five correlated factor model fit better than an independent factor model, but still not adequately

(CFI = .82, TLI = .81, RMSEA = .072), in an Italian community sample. Their results were replicated by Cid and Finney (2009; CFI = .84, RMSEA = .07) in a sample of US undergraduate students.

More recent studies on different populations have also found evidence of the poor fit of a CFA 5-factor model: in Leung, Wong, Chan, and Lam's (2012) data from 439 Chinese smokers who had received a smoking cessation intervention (CFI = .64, SRMR = .09, RMSEA = .06); in Gurven, von Rueden, Massenkoff, Kaplan, and Lero Vie's (2012) study of 632 self-reported and 430 spouse-reported BFI ratings by Bolivian Tsimanes (CFI = .72, RMSEA = .06 and CFI = .52, RMSEA = .08); and Danu's (2013) data from 356 Indonesian participants (CFI = .77, RMSEA = .09). Interestingly, all these studies tested the fit of alternative versions of the BFI developed after the inspection of modification indices, with poor outcomes.

More generally, the failure of CFAs and SEMs to provide clear support for the FFA based on standard measures, has led authors to consider it an 'Achilles' heel' (Furnham, Guenole, Levine, & Chamorro-Premuzic, 2013) of Big Five measures and to conclude that "this points to serious problems with CFA itself when used to examine personality structure" (McCrae, Zonderman, Costa, Bond, & Paunon, 1996, p. 563), since it highlights "not only the limited specificity of personality structure theory, but also the limitations of confirmatory factor analysis for testing personality structure models" (Church & Burke, 1994, p. 93). Church and Burke (1994) argue that the ICM typically used in CFA studies is too restrictive for personality research, since indicators are likely to have secondary loadings unless researchers resort to using a small number of near-synonyms to infer each factor—which would be inconsistent with the wide conceptual breadth of each FFA factor and would be likely to lead to what Cattell (1978) has called 'bloated specifics'. Marsh et al. (2009) suggest that "many ad hoc strategies used to compensate for the inappropriateness of CFA in psychological research more generally are dubious, counterproductive, misleading, or simply wrong" (p. 472). Essentially, the problem is that the strict requirement of zero cross-loadings in typical CFA often does not fit the data well, but leads to a tendency to rely on extensive model modifications to find a well-fitting model.

Allowing some secondary loadings to be non-zero seemed to be the most successful data analytic strategy, as shown, for example, by the aforementioned results of Benet-Martinez and John (1998). However, this strategy appears to be undermined by arbitrariness in the choice of cross-loadings to have free estimation. If items are hypothesized to be complex and to measure multiple aspects of the construct under study, such paths can be specified *a priori*. However, in some cases, there may be no theoretical rationale to inform the analyst's choice of cross-loading to be freed. In such a situation the analyst might revert to using modification indices for exploring and specifying a well-fitting measurement model. As the process of freeing parameters following modification indices is data-driven, the analyst is more susceptible to capitalization on chance characteristics of the data, thus jeopardizing the generalizability of results (e.g., MacCallum, Roznowski, & Necowitz, 1992). Furthermore, misspecification of zero loadings usually leads to distorted factors, with over-estimated factor correlations and subsequent distorted structural relations (Asparouhov & Muthén, 2009; Hopwood & Donnellan, 2010; Marsh et al., 2010).

Some approaches, such as Semi-Confirmatory Factor Analysis (McDonald, 2005), or Partial Confirmatory Factor Analysis (Gignac, 2009) have been proposed to address this issue of taking cross-loadings into account within a confirmatory framework, but they are both confined to examine the factor structure and do not allow for the more sophisticated analyses that are possible in a CFA or a SEM framework: that is, testing multiple group invariance or including observed covariates in the model. Dolan, Oort, Stoel, and Wichterts (2009), foreshadowing the subsequent development of the so-called Exploratory Structural Equation

Modeling (ESEM) approach (Asparouhov & Muthén, 2009), extended the traditional EFA approach based on responses to the NEO-PI-R Big-Five instrument and developed an innovative approach to EFA-based multigroup rotation procedure and tests of measurement invariance (also see Hessen, Dolan, & Wichterts, 2006; Marsh et al., 2010).

ESEM allows both for an EFA method that defines more appropriately the underlying factor structure, and an application of the advanced statistical methods typically associated with CFAs and SEMs. Similarly to an EFA measurement model with rotations, in an ESEM framework all factor loadings are estimated – and thus the ICM-CFA assumption that items must have factorial complexity of one is relaxed. Similarly to CFA and SEM, applied researchers have access to parameter estimates, standard errors, goodness-of-fit statistics, and statistical advances. The only requirement of ESEM is that the number of factors to be extracted has to be specified (For further details of the ESEM approach and identification issues, see Asparouhov & Muthén, 2009)

ESEM has already proved able to overcome the usual shortcomings of using a CFA approach in the study of FFA factors with a variety of Big Five measures (e.g., Cooper et al., 2010; Furnham et al., 2013; Lang, John, Lüdtke, Schupp, & Wagner, 2011; Lavardiére, Morin, & St-Hilaire, 2014; Marsh et al., 2010; Marsh, Nagengast, & Morin, 2012; Samuel, Mullins-Sweatt, & Widiger, 2013) but, to the best of our knowledge, no study has used ESEM to address the issue of the factor structure of the 44-item BFI.

This study is thus a substantive-methodological synergy, testing the usefulness, power, and flexibility of ESEM methods that integrate CFA and EFA to address substantively critical issues about the factor structure of the BFI and its measurement invariance across gender. First, we compare CFA and ESEM approaches, testing whether the assumption that ESEM models fit better than corresponding CFA models is also true in the case of the BFI. Since, in applied settings, unit-weighted sum-of-item scores are routinely used for psychological assessment, we also investigated the overlap between ESEM and CFA factor scores and observed scale scores, as in Furnham et al. (2013). Second, Big Five theory posits that the FFA factors should be substantially orthogonal, but constraining all (non-target) cross-loadings to be zero in the ICM-CFA model typically inflates and biases estimates of factor correlations (Asparouhov & Muthén , 2009). Hence, the support for quasi-orthogonality of BFI factors is hypothesized to be stronger in ESEM models than in CFA models. Third, we exploit the flexibility of ESEM for testing a 13-model taxonomy of measurement invariance, testing invariance across gender of BFI factor loadings, factor variances–covariances, item uniquenesses, uniqueness covariances, item intercepts, and latent means across gender (Marsh et al., 2009).

Previous research on gender differences in BFI scores has always focused on observed scores. In the most comprehensive study carried out so far (Schmitt, Realo, Voracek, & Allik, 2008), in which the BFI was administered to more than 17,000 participants from 55 nations it was found that women scored higher than men in neuroticism, extraversion, agreeableness, and conscientiousness, although differences were lower in less fortunate social and economic conditions. The results from national studies only partially replicated these results, although the direction of the difference was always consistent. Gender differences in neuroticism were found in American (Benet-Martinez & John, 1998), Spanish (Benet-Martinez & John, 1998), French (Plaisant et al., 2010), Italian (Ubbiali et al., 2013), Dutch (Denissen et al., 2008), Chinese (Leung et al., 2013) and German (Lehmann, Denissen, Allemand & Penke, 2013) participants. Differences in agreeableness were also found, except in the Chinese study. Support for differences in conscientiousness was found only in the French and Dutch studies, whereas only Lehmann et al. (2013) found differences in extraversion. Some studies (Denissen et al., 2008; Leung et al., 2013; Lehman et al., 2013) even reported higher openness scores for men. However, unless the underlying BFI factors are measuring the same construct in the same way, and the measurements themselves are operating in the same way across gender, manifest mean comparisons are likely to be invalid. To the best of our knowledge, this is the first study to test gender differences on BFI scores with an ESEM measurement invariance approach.

Method

Participants

Participants were recruited all over Italy through a snowball sampling procedure, in which students or colleagues were given the inventory to pass on to members of their families and acquaintances. Students were selected on the basis of fulfilling the requirements for a degree or a postgraduate training course in psychology. The whole group can thus be considered a convenience sample. The total number of participants was 1,386 (61.8% females), with the mean age being 33.12 years (SD= 14.61, first quartile = 21, median = 27, third quartile = 43, range 18–80). Educational level was low (less than high school) in 11.5% of participants, medium (high school) in 59.7% and high (post-secondary education) in 28.8%. Students composed 41.7% of the participants, 17.9% were office workers, 8.8% were professionals, while the remaining participants were almost equally distributed among other occupations.

All participants volunteered to participate after being presented with a detailed description of the procedure, and all were treated in accordance with the *Ethical Principles of Psychologists and Code of Conduct* (American Psychological Association, 2010). In order to be included in the study, participants had to be at least 18 years-old and to report never having been diagnosed with a psychiatric disorder. Compensation for participation was not given.

Measure

The Italian version of the BFI (Ubbiali et al., 2013, http://www.ocf.berkeley.edu/~johnlab/pdfs/BFI-Italian.pdf) has proven to be reliable, with respect both to internal consistency (Cronbach's αs ranging from .69 to .83) and to temporal stability (test-retest coefficients ranged from .79 to .97). Likewise it has proven to be valid, since scores showed the expected pattern of correlation with scores of the Big Five Questionnaire (Caprara, Barbaranelli, Borgogni, & Perugini, 1993): convergent validity correlations ranged from .56 to .60, discriminant validity correlations from −.21 to .18. Cronbach's αs in this study were .817 (Extraversion; 95% confidence interval [CI]: .802–.831), .693 (Agreeableness; 95% CI: .668–.716), .835 (Conscientiousness; 95% CI: .821–.847), .800 (Neuroticism; 95% CI: .784–.816), and .810 (Openness; 95% CI: .795–.825).

Total Group Analyses

Analyses were conducted with Mplus 6© (Muthén & Muthén, 1998–2010). Preliminary analyses consisted of a traditional CFA on the total group of participants, based on the Mplus robust maximum likelihood estimator (MLR), with standard errors and tests of fit that were robust in relation to the nonnormality of observations (Muthén & Muthén, 1998–2010). Then the ESEM was applied to responses to the BFI. We used an oblique GEOMIN rotation (the default in Mplus) with an epsilon value of .5.

Measurement Invariance Models

Measurement invariance across gender was tested through a 13-model taxonomy of invariance tests that integrated factor and measurement invariance traditions (for a more detailed discussion of the invariance models see Marsh et al., 2009, 2010). Following Meredith (1993), the sequence of invariance testing begins with a model of 'configural' invariance: that is, with no invariance of any parameter estimates (i.e., all parameters are freely estimated), such that only the similarity of the overall pattern of parameters is evaluated. Since it does not require any estimated parameters to be the same, this model is not an actual invariance model. However, its fit must be evaluated: First, the ability of the a priori model to fit the data in each group without invariance constraints must be tested. Second, a baseline for comparing other models that impose equality constraints on the parameter estimates across groups can be provided.

The next step in invariance testing is to test a ‘weak’ measurement invariance model; this requires that factor loadings are invariant over groups. If indicator means (i.e., the intercepts of responses to individual items) are also constrained to be equal across groups, then a ‘strong’ measurement invariance model is specified. If such a model fits, factor loadings and item intercepts are invariant over groups, and changes in the latent factor means can reasonably be interpreted as changes in the latent constructs, since they have been corrected for measurement error. A power analysis (Muthén & Muthén, 2002) showed that the sample at hand afforded sufficient statistical power (i.e., .80) to test this model (details of the analysis are available upon request from the corresponding author).

Further, factor loading and item intercept invariance is a necessary but not sufficient condition for testing *manifest* group mean differences, which also require invariance of item uniquenesses. The presence of differences in reliability (as represented or absorbed in the item uniquenesses) across groups could in fact distort mean differences on the observed scores. A model that specifies the invariance of item uniquenesses is referred to as a ‘strict’ invariance model.

Recently, Marsh et al. (2009) expanded this measurement invariance tradition, suggesting a taxonomy of 13 partially nested models, with models varying from the least restrictive model of configural invariance to a model of complete invariance that posits strict invariance, together with invariance of the latent means and of the factor variance-covariance matrix (see below Table 3 for a description). The invariance of the factor variance-covariance matrix is not a prior focus of measurement invariance, as it does not compromise comparisons of latent mean differences across groups. However, it is often crucial in studies on the invariance of covariance structures of multifactorial constructs, like the BFI factors, since the pattern of relations among factors might have important practical and/or theoretical implications. (see Marsh et al., 2009 for a more extended discussion of these issues)

Typically, models of measurement invariance are tested within a CFA framework. In this study we used tests of measurement invariance over gender on the basis of a similar taxonomy of invariance tests, but within an ESEM framework.

Correlated Uniquenesses

For both CFA and ESEM models, we included both freely estimated uniquenesses (reflecting a combination of measurement-error-specific variances) and a priori correlated uniquenesses (CUs; covariances between the specific variance components associated with two different items from the same FFA factor). In general, using ex post facto CUs should be avoided (e.g., Marsh, 2007), although there are some circumstances in which a priori CUs should be specified (e.g., when the same items are used on multiple occasions, since the correlation of unique components of the same item administered on different occasions cannot be explained simply in terms of correlations between the factors).

However, an increase in model fit due to freeing error covariances is usually the result of further shared variance among items, other than that explained by the specified latent factors. This may result from method effects (such as in the common measurement method of self-report), from similar wording of items (e.g., positive or negative phrasing) or from ‘specific’ or ‘group’ factors that are independent of the ‘general’ factor (e.g., Brown, 2006). Since the emergence of FFA, it has been pointed out that describing personality in terms of five broad domains can be efficient, as it allows for the prediction of many outcomes, with modest to moderate levels of precision (e.g., John, Hampson, & Goldberg 1991).

However, a crucial limitation of investigating personality in terms of the five broad domains is their low fidelity (Soto & John, 2009). Each domain subsumes more specific personality characteristics, sometimes referred to as 'facets' (e.g., Costa & McCrae, 1992). Merging these related but distinguishable facet traits into broad domains results in a loss of information, thus reducing a scale's ability to describe, predict, and explain behavior. This so-called 'bandwidth-fidelity dilemma' (Cronbach & Gleser, 1957) can be resolved in a hierarchical model of personality. The measures developed in such a framework allow for the assessment both of the five broad domains and of more specific traits within those domains (e.g., NEO PI-R).

Nevertheless, the use of such measures in research studies is usually limited by the fact that they include hundreds of items and require a long administration time. Hence, shorter measures, like the BFI or the NEO-FFI, which provide scores only in the five broad domains, are often preferred. However, in the development of the BFI, the selection of items to best represent each of the Big Five factors was made without explicit reference to the facets. Thus, some facets may be overrepresented, whereas others may be represented by a single item or else not represented at all.

This issue is not new in research on FFA measures: Marsh et al. (2010) noted that in the construction of the NEO-FFI, items were selected from the whole NEO-PI-R pool to best represent each of the Big Five factors, solely on the basis of their correlation with the factor score at the domain level, and without reference to the facets. Marsh et al. (2010) posited that items that coming from the same facet of a specific Big Five factor would have higher correlations (that is, beyond those that could be explained in terms of the common Big Five factor that they represented) than would items that came from different facets of the same Big

Five factor. They thus decided to model these potentially inflated correlations that were due to facets, as CUs, relating each pair of items from the same facet.

In the case of the BFI, based on content and correlational analyses, Soto and John (2009) identified from the item pool of the BFI, ten specific facet traits that converged with the NEO-PI-R facets and that also corresponded well with lower-level traits identified by other hierarchical Big Five models. These results were consistent with Chiorri et al.'s (2008) modification indices, which suggested to free the correlation among error variances of items subsequently indicated by Soto and John (2009) as belonging to the same facet.

Following Marsh et al. (2010) we thus decided to specify CUs for these items. This resulted in an a priori set of 61 CUs inherent to the design of the BFI (note that Soto and John included only 35 out of the original 44 items in the facets). Moreover, based on the results of Chiorri et al. (2008) we decided to model two further CUs (between Items 26 and 11, and between 43 and 8), which were not suggested by the facets identified by Soto and John (2009) but which seemed to contribute to an adequate model fit, apparently due to a wording effect that was idiosyncratic to the Italian translation of the BFI items. Although we argue that this set of a priori CUs should be included in all factor analyses of (Italian) BFI responses, we systematically evaluated models with and without these CUs, as well as the invariance of these CUs over gender.

Goodness of fit

Although no study has yet focused on the appropriateness of the traditional CFA indices of fit for ESEM, we followed previous studies (e.g., Marsh et al., 2010) in considering the comparative fit index (CFI), the Tucker–Lewis index (TLI), the root-mean-square error of approximation (RMSEA) and the significance of parameter estimates. For both the CFI and TLI, values greater than .90 and .95 are considered to reflect acceptable and optimal fits respectively, to the data. For the RMSEA, values less than .08 and .06 are considered respectively as indices of reasonable and optimal fit to the data (Marsh, Hau, & Wen, 2004).

In the comparison of nested invariance models, we considered support for the more parsimonious model to be evidenced in a change in the CFI of less than .01 (Chen, 2007; Cheung & Rensvold, 2002) or a change in the RMSEA of less than .015 (Chen, 2007). Since, as noted by Marsh (2007), some indices (e.g., TLI and RMSEA) incorporate a penalty for parsimony, so that the more parsimonious model can fit the data better than can a less parsimonious model (i.e., the gain in parsimony is greater than the loss in fit), we also considered as support for the more parsimonious model a TLI or RMSEA which was as good as or better than, that for the more complex model.

Results

Total Group Analyses

We first tested the hypothesis that the ESEM model provides a better fit to BFI responses than does a traditional ICM-CFA model. As in previous studies of Big Five measures, ESEM performed noticeably better than the ICM-CFA model in terms of goodness of fit. The ICM-CFA model, that did incorporate the 63 a priori CUs based on the facet structure of the BFI, did not provide an acceptable fit to the data ($X^2(892) = 5879.232$, CFI = .700, TLI = .682, RMSEA = .064). The fit of the model that specified the a priori CUs was still inadequate, although improved ($X^2(829) = 3715.740$, CFI = .826, TLI = .802, RMSEA = .050). The corresponding ESEM solutions fitted the data much better. The fit of the total group ESEM with no a priori CUs was not acceptable ($X^2(736) = 3415.029$, CFI = .839, TLI = .793, RMSEA = .051), whereas the inclusion of CUs allowed the model to reach an acceptable fit ($X^2(673) = 1823.265$, CFI = .931, TLI = .903; RMSEA = .035). However, ESEM is an *exploratory* method, as is EFA: hence, one needs to examine and interpret the patterns of factor loadings, and the significance of the loadings for each of the factors. In fact, it is possible that the pattern of factor loadings will not support the structure specified in the corresponding CFA model. In such a case we would expect that, based on their highest significant loading, items could be grouped into their expected factors. Nonetheless, given the results of previous studies that applied CFA to BFI data (e.g., Benet-Martinez & John, 1998), we also expected that some cross-loadings, without being higher than the target loadings, should be significantly different from zero.

As shown in Table 1, the expected pattern of results appears to have been supported.

[Table 1]

In both types of models, factor loadings tended to be modest, with few loadings greater than .70 (7 in CFA, 5 in ESEM models) and almost no target loading less than .30 (2 in the CFA, 1 in the ESEM solution). CFA factor loadings (Median [Mdn] = .54) were similar to target loadings in the ESEM model (Mdn = .51), and the pattern of loadings was nearly identical for the two solutions.

To provide an objective evaluation of these results, we computed a profile similarity index (PSI) correlating the vector of 44 CFA factor loadings with the corresponding vector of 44 ESEM target loadings. The PSI suggested that ESEM and CFA target factor loadings were highly related (r = .850, 95% CI: .740–.916).

As for cross-loadings in the ESEM solution, more than half (99 out of 176) were statistically different from zero, ranging from -.25 to .33. However, only one cross-loading was higher than .30 (Agreeableness, item 42, loaded also on the Extraversion factor) and for no item was a cross-loading higher than the target loading.

Although patterns of correlations (Table 1) were similar, the CFA factor correlations (-.26 to .49, Mdn absolute value = .24) tended to be systematically larger than those for ESEM (-.15 to .21; Mdn absolute value = .10). Thus, for example, the positive correlation between

Openness and Extraversion was .49 on the basis of the CFA solution, but only .21 in the ESEM solution. Similarly, the negative correlations of Neuroticism with Extraversion, Agreeableness and Conscientiousness were in the .20s in the CFA solution, but not higher than .10 in the ESEM solution.

Since, in applied assessment settings, unit-weighted sum-of-item scores are routinely used, we computed the correlations of the ESEM and CFA factor scores with observed scale scores. The results showed that the correlations between both forms of latent score (i.e., ESEM and CFA factor scores) and observed scale scores were statistically equal to or higher than .90 in all cases (see Table 2 below).

Invariance Over Gender

Usually, gender differences are tested through comparisons on raw scores, not corrected for measurement error. Based on the observed scores reported in Table 2, and applying the Benjamini and Hochberg (1995) step-up false discovery rate-controlling procedure for controlling the inflation of Type I Error due to multiple comparisons, we would draw the conclusion that females in this study scored significantly higher than men on Neuroticism (Difference 95% CI: 0.305–0.468, $t(1384) = 9.28$, adjusted $p < .001$, $r = .242$), and Agreeableness (0.052–0.181, $t(1384) = 3.56$, adjusted $p = .001$, $r = .095$), but did not differ from males in Conscientiousness (0.005–0.165, $t(1384) = 2.08$, adjusted $p = .062$, $r = .056$), Extraversion (-0.005–0.158, $t(1384) = 1.84$, adjusted $p = .082$, $r = .049$) and Openness (-0.113–0.034, $t(1384) = -1.05$, adjusted $p = .294$, $r = .028$).

[Table 2]

However, unless factor loadings, item intercepts, and uniquenesses are shown to be invariant across gender, such comparisons as the above are likely to be invalid. To address this issue, we applied Marsh et al.'s (2009) taxonomy of 13 ESEM models. In the present study the application of this taxonomy of models is complicated by the CUs, which are necessary to achieve an acceptable fit to the data in the total group. Hence, it was also needed to determine the extent to which these CUs were invariant over gender and how this influenced the results of the various models. For all 13 models we first tested models with no CUs (e.g., MG1 in Table 3 corresponds to the first model in the invariance taxonomy), and then tested two additional models, one in which the CUs were allowed to vary for females and males (submodels labeled A in the Description column of Table 3, as in ‘MG1A’) and another in which the CUs were constrained to be invariant over responses by females and males (submodels labeled B in Table 3, as in ‘MG1B’).

[Table 3]

As a result, within this set of three submodels there was a systematic nesting to evaluate the CUs and their invariance across gender in relation to each of the 13 invariance models.

Model MG1, with no invariance constraints, did not provide an acceptable fit to the data (CFI = .829, TLI = .780). These fit statistics are similar to those based on the total group ESEM model. However, consistent with earlier results, the inclusion of the set of a priori CUs substantially improved the fit to a marginally acceptable level (CFI = .923, TLI = .892; see MG1A in Table 3). Importantly, constraining these a priori CUs to be invariant over gender (see MG1B in Table 3) resulted in nearly no change in fit (CFI = .921, TLI = .894). For fit indices controlling for parsimony, the fit was substantially unchanged or slightly better for MG1B than for MG1A (.892 to .894 for TLI; .037 to .037 for RMSEA). For the CFI, which is monotonic with parsimony, the change (.923 to .921) was clearly less than the .01 value usually considered to be in support of invariance constraints.

These results, demonstrating that the sizes of the 63 CUs are reasonably invariant over gender, are substantively important. For each of the 13 models used to test the factorial invariance of the full mean structure, the inclusion of such a set of CUs noticeably improved the goodness of fit. The results of a comparison of the models in which CUs were freely estimated against those in which CUs were constrained to be equal across gender, support the invariance of the CUs. Hence, the high consistency of this pattern of results over the different models, provides clear support for the inclusion of these CUs in the design of the BFI. Thus, the presentation of results focuses on the models, including gender-invariant CUs (e.g., Model MG1B for Model 1). Factor loadings, uniquenesses, intercepts, and factor correlations and their standard errors for the configural invariance model are available upon request from the corresponding author.

Weak factorial/measurement invariance model. Model MG2B (along with MG2 and MG2A in Table 3) tested the invariance of factor loadings over gender: that is, whether the factor loadings are the same for females and males. The critical comparison between the more parsimonious MG2B (with factor loadings invariant) and the less parsimonious MG1B (with no factor loading invariance) supports the invariance of factor loadings over gender. Fit indices that control for model parsimony are as good or better for the more parsimonious MG2B (TLI = .902 vs. .894; RMSEA = .035 vs. .037), whereas the difference in CFI (.917 vs. .921) is less than the value of .01, which typically is used to reject the more parsimonious model.

Strong measurement invariance model. This model requires that item intercepts, along with factor loadings, be invariant over groups. The critical comparison is thus between Models MG2B and MG5B: that is, whether differences in the 44 intercepts can be explained in terms of five latent means (i.e., a complete absence of differential functioning). The fit of MG5B (CFI = .913, TLI = .900, RMSEA = .036) can be considered equal to the fit of the corresponding model MG2B (CFI = .917, TLI = .902; RMSEA = .035). These results suggest that item intercepts are invariant, that gender differences at the level of item means can be explained in terms of the factor means, and that there is no differential item functioning between gender groups.

Strict measurement invariance model. This model requires that item uniquenesses, item intercepts, and factor loadings all be invariant over the groups. The crucial comparison is between Models MG5B and MG7B. Models MG5B and MG7B showed a very similar fit to the data (CFI = .913 vs .909; TLI = .900 vs. .898; RMSEA = .036 vs. .036). Furthermore, a comparison of all the other various pairs of models that tested the invariance of the uniquenesses (MG3B vs. MG2B; MG6B vs. MG4B; MG9B vs.MG8B; MG11B vs. MG10B; MG13B vs. MG12B) yielded the same results. Hence, it can be concluded that BFI item uniquenesses are invariant over gender.

Factor variance–covariance invariance. This model requires that the variance-covariance matrices of the BFI factors be invariant over the groups. The crucial comparison is between Models MG2B (factor loadings invariant) and MG4B (factor loadings and factor variance–covariance invariant). The results provided support for the additional invariance constraints, in terms of the values of the fit indices (CFI = .917 vs .917; TLI = .903 vs. .902, RMSEA = .035 vs .035) and of their comparison with MG2. Further tests of the invariance of the latent factor variance-covariance matrix could be based on any pair from the six models in Table 3. The items in each pair differ only in relation to whether the factor variance-covariance matrix is free, or not (i.e., MG6B vs MG3B, MG8B vs MG5B, MG9B vs MG7B, MG12B vs MG510B, MG13B vs MG11B). Note that if there were systematic and substantive differences in the interpretations on the basis of these different comparisons, true differences in the factor variance-covariance matrix could conceivably be ‘absorbed’ into differences in other parameters that had not been constrained to be invariant. However, this complication does not seem to have been the case in the present investigation, since support for the invariance of the factor variance–covariance matrix is consistent across each of these alternative comparisons.

Latent factor means invariance. The last four models (see MG10–MG13 in Table 3) in the taxonomy all constrain mean differences between men and women to be zero—in combination with the invariance of other parameters. The critical comparisons for testing gender mean invariance are MG10B vs MG5B, MG11B vs MG7B, MG12B vs MG8B, and MG13B vs MG9B. In all these cases the differences in fit indices supported invariance only marginally (e.g., model MG10 vs MG5: CFI = .905 vs .913; TLI = .891 vs. .900). Modification indices for model MG10B suggested freeing latent means for all factors except Openness, implying that on this dimension, scores do not differ systematically for females and males. Unfortunately, Mplus does not allow for testing partial invariance of factor means in ESEM, since they must have the same constraints. However, examining models in which means were constrained to 0 in one group (females) and freely estimated in the other group (males), it was apparent that females yielded significantly higher scores on neuroticism, agreeableness, extraversion and conscientiousness, whereas the difference in Openness latent means was not significant. Standardized gender differences on the basis of each of the 12 models that provided estimates of these differences, are summarized in Table 2.

Discussion

The present study is a substantive-methodological synergy, applying a new and evolving methodological innovation, Exploratory Structural Equation Modeling (ESEM) to explore some psychometric properties (factor structure, correlations among factors and measurement invariance across gender) of the Big Five Inventory (BFI). In recent years, a few studies (Marsh et al., 2009, 2010, 2012) have argued that the traditional ICM–CFA model is not appropriate for many well-established psychological measures, including most FFA measures, and further, that this position has been shared by FFA researchers for years (e.g., Church & Burke, 1994; McCrae et al., 1996).

As noted in the introduction, research on the BFI factor structure has almost always been undertaken with the use of varimax-rotated PCA. Though there are possible reasons for doing this (e.g., computing congruence coefficients to compare solutions from different adaptation studies, as in Ubbiali et al., 2013), there seems to be no reason to avoid a confirmatory approach, other than obtaining poor goodness of fit indices. This study aimed to address this issue through the application of ESEM, and the expected BFI five-factor structure was found using ICM-CFA and ESEM. The pattern, and even the size of target factor loadings, was similar for the two approaches.

However, the ESEM solution showed that more than half of the cross-loadings were statistically different from zero: It is then not surprising that the ICM-CFA solution, which constrained these loadings to zero, had a substantially worse fit, as this is consistent with previous results on the BFI and other FFA measures. Furthermore, the factor score estimates based on the ESEM model correlated almost perfectly both with the scores estimated on the basis of the CFA model and on their unit-weighted sum-of-item score counterparts, suggesting that the observed scale scores routinely used are appropriate for the assessment of personality trait levels, as measured by the BFI.

Another advantage of ESEM is its ability to address issues related to complex structures of measurement error in CFA, overcoming both the lack of definition and the lack of control for measurement error in traditional EFA approaches, and the need for constraints on factor loadings imposed in the traditional ICM–CFA approach. The commonly reported internal consistency estimates of reliability ignore other aspects of unreliability, and do not correct parameter estimates for it (see Sijtsma, 2009). Further, the failure to control for complex structure of measurement error can have unanticipated results (see discussion of the ‘phantom effect’ by Marsh et al., 2010).

In the present study, ESEM allowed us to model an additional source of measurement error that could be idiosyncratic to the design of FFA shorter measures such as the BFI (see Introduction). That is, we posited that items that had been identified by previous research (Soto & John, 2009) as belonging to the same facet would be more highly correlated than would items from different facets designed to measure the same factor. Consistently with previous empirical findings (e.g., Chiorri et al., 2008), we found support for this additional source of measurement error, since inclusion of CUs contributed substantially to goodness of fit, and the CUs were invariant over responses by men and women. Although these CUs are idiosyncratic to the design of the BFI, it was possible that other method effects, such as wording effects, could distort the findings if not controlled for. Accordingly, we tested alternative models in which specific wording factors were specified, but their fit was only marginally acceptable, and was worse than models with CUs (details of these analyses are available upon request from the corresponding author).

The ESEM solution also resulted in substantially less correlated factors than did CFA. This result is consistent with previous results employing the same methodology on other FFA measures (e.g., Cooper et al., 2010; Furnham et al., 2013; Lang et al., 2011; Lavardiére et al., 2014; Marsh et al., 2010, 2012; Samuel et al., 2013), and with the Big Five theory itself, which assumes (quasi) orthogonality among factors. In an ICM-CFA solution, the relation between a specific item and a nontarget factor that would be accounted for by a cross-loading can be represented only through the factor correlation between the two factors. If there are at least moderate cross-loadings in the true population model and these are constrained to be zero, as in the ICM-CFA model, then estimated factor correlations are likely to be inflated (e.g., Asparouhov & Muthén, 2009). This could result in multicollinearity and undermine discriminant validity, in relation to predicting other outcomes and providing distinct profiles of personality. Moreover, Ashton, Lee, Goldberg, and De Vries (2009) argue that higher order personality factors accounting for these correlations will be spurious, because the correlations on which they are based are artifactual.

In this study we pursued issues in latent BFI factors with appropriate tests of full measurement and structural invariance, in relation to a comprehensive taxonomy of invariance models. Multi-group ESEM analyses supported invariance over gender of factor loadings, item intercepts and uniquenesses, correlated uniquenesses and factor variances and covariances. These analyses could not have been performed appropriately with traditional EFA approaches or with ICM-CFA models that were not able to fit the data. Whereas observed score comparisons were significant only for neuroticism, agreeableness and conscientiousness, measurement invariance analyses revealed that, consistently with previous research based on BFI manifest scores (e.g., Schmitt et al., 2008) and with other studies employing the same method on FFA measures (e.g., Marsh et al., 2010), women scored higher on all five BFI factors except Openness. This provides more reliable evidence of gender differences in BFI scores. Gender differences in personality traits can be explained from a variety of perspectives: biological, evolutionary, biosocial, sociocultural, etc.; these are reviewed thoroughly in Schmitt et al. (2008).

The major limitation of this study is the reliance on a convenience sample, which limits the external validity and the generalizability of the results, and does not rule out capitalization on chance, given the risk of biases due to the recruitment procedure. Although the relatively large sample size and the variety of geographical regions and socio-economic backgrounds from which the participants came may well have, in their turn, limited possible biases, we could not address another major issue in research on personality—namely, age effects.

Recent research (e.g., Marsh et al., 2012; Wortman, Lucas, & Donnellan, 2012) has relied on very large and nationally representative samples that allowed a reliable estimation of age effects across the whole range of age scores. Although these studies showed intriguing linear and non-linear effects, detectable through the availability of participants older than 60 (e.g., Marsh et al., 2012), given that that only 86 participants (6.1%) in this study were older than 60, any analysis of age effects would be inconclusive and not comparable to the latest evidence.

It must also be noted that, while we support the ESEM model as a viable alternative to the traditional ICM–CFA model, we do not intend to suggest that the ESEM approach should in all cases replace the CFA approach. One shortcoming of ESEM is that it is less parsimonious than a corresponding ICM-CFA model: hence, when a ICM–CFA model fits the data as well as the ESEM model does, and results in similar parameter estimates, the ICM–CFA should be used. Where the ICM–CFA does not provide an adequate fit to the data (and therefore the assumptions of the ICM–CFA model are unlikely to be valid) but the ESEM model does, we do suggest that advanced statistical strategies such as those presented here are more appropriately conducted with ESEM models than with ICM–CFA models, since the less restrictive assumptions of the ESEM model provide more valid parameter estimates. Finally, the pattern of factor loadings, and its significance, must be examined, to check whether it matches theoretical expectations.

These limitations aside, this study has provided support for the five-factor structure of the (Italian) BFI, the adequacy of its unit-weighted sum-of-item scores, and its measurement invariance across gender, using the ESEM: a relatively new, methodologically sound and flexible modeling approach that allows for addressing issues for which the traditional EFA and ICM-CFA approaches are not well-suited.

References

American Psychological Association (2010). Ethical principles of psychologists and code of conduct. Retrieved from http://www.apa.org/ethics/code/principles.pdf

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.

Benet-Martínez, V., & John, O. P. (1998). Los Cinco Grandes across cultures and ethnic groups: multitrait multimethod analyses of the Big Five in Spanish and English. *Journal of Personality and Social Psychology*, *75*, 729–750.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B, 57*, 289–300.

Block, J. (2010). The five-factor framing of personality and beyond: Some ruminations. *Psychological Inquiry*, *21*, 2–25.

Bollen, K. A., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin*, *110*, 305–314.

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

Caprara, G. V., Barbaranelli, C., Borgogni, L., & Perugini, M. (1993). The Big 5 Questionnaire: A new questionnaire to assess the 5 factor model. *Personality and Individual Differences*, *15*, 281–188.

Cattell, R. B. (1978). *Scientific Use of Factor Analysis in Behavioral and Life Sciences*. New York, NY: Plenum Press.

Chen, F. F. (2007). Sensitivity of goodness of fit indices to lack of measurement invariance. *Structural Equation Modeling, 14*, 464–504.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indices for testing measurement invariance. *Structural Equation Modeling, 9*, 233–255.

Chiorri, C., Ubbiali, A., & Donati, D. (2008, June). *Can personality traits be reliably assessed with short measures? An Italian study on the shortened version of the Big Five Inventory-44*. Paper presented at the 39th International Meeting of the Society for Research in Psychotherapy, Barcelona, Spain, June 18–21. Retrieved from http://www.slideshare.net/carloskij/bfibarcelona08

Church, A. T., & Burke, P. J. (1994). Exploratory and confirmatory tests of the Big Five and Tellegen's three- and 4-dimensional models. *Journal of Personality and Social Psychology, 66*, 93–114.

Cid, J., & Finney, S. J. (2009, April). *Confirmatory factor analysis of the Big Five Inventory*. Paper presented at the 2009 American Educational Research Association Annual Meeting, San Diego, CA, April 13–17.

Cooper, A., Smillie, L. D., & Corr, P. J. (2010). A confirmatory factor analysis of the Mini-IPIP Five-Factor Model personality scale. *Personality and Individual Differences, 48*, 688–691.

Costa, P. T. Jr., & McCrae, R. R. (1992). *NEO-PI-R professional manual*. Odessa, FL: Psychological Assessment Resources.

Cronbach, L. J., & Gleser, G. C. (1957). *Psychological tests and personnel decisions*. Urbana, IL: University of Illinois Press.

Danu, D. A. (2013). *Uji Validitas Konstrukt Big Five Inventory dengan Pendekatan Analisis Faktor Konfirmatori. Ringkasan Skripsi. Tidak diterbitkan* [Testing the construct validity of the Big Five Inventory through a confirmatory factor analysis approach. Unpublished dissertation]. Yogyakarta: Fakultas Psikologi Universitas Gadjah Mada. Retrieved from https://id.scribd.com/doc/133344882/Uji-Validitas-Konstruk-Big-Five-Inventory-dengan-Pendekatan-Analisis-Faktor-Konfirmatori

Denissen, J. J. A., Geenen, R., van Aken, M. A. G., Gosling, S. D., & Potter, J. (2008). Development and validation of a Dutch translation of the Big Five Inventory (BFI), *Journal of Personality Assessment, 90*, 152–157

Dolan, C. V., Oort, F. J., Stoel, R. D., & Wichterts, J. M. (2009). Testing measurement invariance in the target rotated multigroup exploratory factor model. *Structural Equation Modeling, 16*, 295–314.

Furnham, A., Guenole, N., Levine, S. Z., & Chamorro-Premuzic, T. (2013). The NEO Personality Inventory-Revised: Factor structure and gender invariance from exploratory structural equation modeling analyses in a high-stakes setting. *Assessment*, *20*(1), 14–23.

Gignac, G. E. (2009). Partial confirmatory factor analysis: described and illustrated on the NEO-PI-R. *Journal of Personality Assessment, 91*, 40–47.

Gurven, M., von Rueden, C., Massenkoff, M., Kaplan, H., & Lero Vie, M. (2013). How universal is the Big Five? Testing the five-factor model of personality variation among forager–farmers in the Bolivian Amazon. *Journal of Personality and Social Psychology, 104,* 354–370.

Hessen, D. J., Dolan, C. V., & Wicherts, J. M. (2006). The multi-group common factor model with minimal uniqueness constraints and the power to detect uniform bias. *Applied Psychological Measurement, 30,* 233–246.

Hopwood, C. J. & Donnellan, M. B. (2010). How should the internal structure of personality inventories be evaluated? *Personality and Social Psychology Review, 14,* 332–346.

John, O. P., Donahue, E. M., & Kentle, R. L. (1991). *The Big Five Inventory-Versions 4a and 54*. Berkeley, CA: University of California, Berkeley, Institute of Personality and Social Research.

John, O. P., Hampson, S. E., & Goldberg, L. R. (1991). The basic level of personality-trait hierarchies: Studies of trait use and accessibility in different contexts. *Journal of Personality and Social Psychology, 60,* 348–361.

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative Big Five trait taxonomy: History, measurement, and conceptual issues. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (3rd ed., pp. 114–158). New York, NY: Guilford.

Lang, F. R., John, D., Lüdtke, O., Schupp, J., & Wagner, G. G. (2011). Short assessment of the Big Five: robust across survey methods except telephone interviewing. *Behavior Research Methods, 43*, 548–567.

Lang, F., Lüdtke, O. & Asendorpf, J. (2001). Testgüte und psychometrische Äquivalenz der deutschen Version des Big Five Inventory (BFI) bei jungen, mittelalten und alten Erwachsenen [Adequacy and psychometric equivalence of the German version of the Big Five Inventory (BFI) in young, middle-aged, and old adults]. *Diagnostica ,47*, 111–121.

Lavardiére, O., Morin, A. J. S., & St-Hilaire, F. (2014). Factor structure and measurement invariance of a short measure of the Big Five personality traits. *Personality and Individual Differences, 55,* 739–743.

Lehmann, R., Denissen, J. J. A., Allemand, M., & Penke, L. (2013). Age and gender differences in motivational manifestations of the Big Five from age 16 to 60. *Developmental Psychology, 49*, 365–383.

Leung, D. Y., Wong, E. M., Chan, S. S., & Lam, T. H. (2013). Psychometric properties of the Big Five Inventory in a Chinese sample of smokers receiving cessation treatment: A validation study. *Journal of Nursing Education and Practice, 3*(6), doi: 10.5430/jnep.v3n6p1.

Levine, S. Z., & Jackson, C. J. (2002). Aggregated personality, climate and demographic factors as predictors of departmental shrinkage. *Journal of Business and Psychology, 17,* 287–297.

Little, T. D., Rhemtulla, M., Gibson, K., & Schoemann, A. M. (2013). Why the items versus parcels controversy needn't be one. *Psychological Methods, 18,* 285–300.

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin, 111,* 490–504.

Marsh, H. W. (2007). Application of confirmatory factor analysis and structural equation modeling in sport/exercise psychology. In G. Tenenbaum & R. C. Eklund (Eds.), *Handbook of sport psychology* (3rd ed., pp. 774–798). Hoboken, NJ: Wiley.

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

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.

Marsh, H. W., Lüdtke, O., Nagengast, B., Morin, A. J. S., & Von Davier, M. (2013). Why item parcels are (almost) never appropriate: Two wrongs do not make a right—camouflaging misspecification with item parcels in CFA models. *Psychological Methods*, *18,* 257–284.

Marsh, H. W., Muthén, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J. S., & Trautwein, U. (2009). Exploratory structural equation modeling, integrating CFA and

EFA: Application to students' evaluations of university teaching. *Structural Equation Modeling, 16*, 439–476.

Marsh, H. W., Nagengast, B., & Morin, A. J. (2012). Measurement invariance of big-five factors over the life span: ESEM tests of gender, age, plasticity, maturity, and la dolce vita effects. *Developmental Psychology, 49*, 1194–1218.

McCrae, R. R., & Costa, P. T., Jr. (1997). Personality trait structure as a human universal. *American Psychologist, 52*, 509–516.

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

McDonald, R. P. (2005). Semiconfirmatory factory analysis: The example of anxiety and depression. *Structural Equation Modeling, 12*, 163–172.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance, *Psychometrika, 58*, 525–543.

Muthén, B. & Muthén, L. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling, 4*, 599–620.

Muthén, L. K., & Muthén, B. (1998–2010). *Mplus user's guide*. Los Angeles, CA: Muthén & Muthén.

Plaisant, O., Courtois, R., Réveillère, C., Mendelsohn, G. A., & John, O. P. (2010). Validation par analyse factorielle du Big Five Inventory français (BFI-Fr). Analyse convergente avec le NEO-PI-R [Validation through factor analysis of the French Big Five Inventory (BFI-Fr). Convergent analysis with the NEO-PI-R]. *Annales Médico-Psychologiques, 168*, 97–106.

Samuel, D. B., Mullins-Sweatt, S. N., & Widiger, T. A. (2013). An investigation of the factor structure and convergent and discriminant validity of the Five-Factor Model Rating Form. *Assessment*, *20*, 24–35.

Schmitt, D. P., Allik, J., McCrae, R. R., & Benet-Martínez, V. (2007). The geographic distribution of Big Five personality traits: Pattern and profiles of human self-description across 56 nations. *Journal of Cross-Cultural Psychology*, *38*, 173–212.

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*, 168–182.

Sijtsma, K. (2009). Reliability beyond theory and into practice. *Psychometrika, 74*, 169–173.

Soto, C. J., & John, O. P. (2009). Ten facet scales for the Big Five Inventory: Convergence with NEO PI-R facets, self-peer agreement, and discriminant validity. *Journal of Research in Personality, 43*, 84–90.

Ubbiali, A., Chiorri, C., Hampton, P., & Donati, D. (2013). Psychometric properties of the Italian adaptation of the Big Five Inventory (BFI). *Bollettino di Psicologia Applicata, 266*, 37–46.

Vandenberghe, C., St-Onge, S., & Robineau, É. (2008) An analysis of the relation between personality and the attractiveness of total rewards components. *Relations Industrielles/Industrial Relations, 63*, 425–453.

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

Velicer, W. F., & Jackson, D. N. (1990). Component analysis vs. common factor analysis: Some further observations. *Multivariate Behavioral Research, 25*, 95–112.

Wortman, J., Lucas, R. E., & Donnellan, M. B. (2012). Stability and change in the big five personality domains: evidence from a longitudinal study of Australians. *Psychology and Aging*, *27*, 867–874.

Table 1 *Confirmatory Factor Analysis and Exploratory Structural Equation Modeling Standardized Factor Loadings, Uniquenesses and Factor Correlations Based on Responses to the Big Five Inventory (n = 1,386)*

	CFA solution						ESEM solution					
Item	E	A	C	N	O	Uniq	E	A	C	N	O	Uniq
1. is talkative	**.65**	.00	.00	.00	.00	**.58**	**.68**	.02	.00	**.09**	**.05**	**.53**
6. is reserved*	**.27**	.00	.00	.00	.00	**.82**	**.48**	-.05	**-.22**	**.09**	**-.10**	**.77**
11. is full of energy	**.53**	.00	.00	.00	.00	**.42**	**.39**	.04	**.25**	**-.19**	**.16**	**.42**
16. generates a lot of enthusiasm	**.64**	.00	.00	.00	.00	**.74**	**.50**	**.09**	.05	**-.08**	**.24**	**.68**
21. tends to be quiet	**.57**	.00	.00	.00	.00	**.51**	**.65**	-.01	-.03	.03	**-.06**	**.50**
26. has an assertive personality*	**.70**	.00	.00	.00	.00	**.93**	**.51**	-.04	**.24**	**-.14**	**.24**	**.74**
31. is sometimes shy, inhibited	**.43**	.00	.00	.00	.00	**.81**	**.45**	**-.13**	.00	**-.24**	-.06	**.65**
36. is outgoing, sociable*	**.76**	.00	.00	.00	.00	**.83**	**.80**	**.10**	-.01	.00	**.04**	**.77**
2. tends to find fault with others*	.00	**.43**	.00	.00	.00	**.48**	**-.13**	**.39**	-.01	**-.25**	-.05	**.58**
7. is helpful and unselfish with others	.00	**.44**	.00	.00	.00	**.68**	**.15**	**.50**	**.13**	**.17**	**.11**	**.69**
12. starts quarrels with others	.00	**.42**	.00	.00	.00	**.72**	**-.18**	**.42**	.01	**-.16**	**-.09**	**.63**
17. has a forgiving nature*	.00	**.47**	.00	.00	.00	**.82**	-.03	**.50**	**-.14**	-.05	.04	**.77**
22. is generally trusting	.00	**.43**	.00	.00	.00	**.46**	-.01	**.45**	-.03	.00	-.05	**.45**
27. can be cold and aloof	.00	**.33**	.00	.00	.00	**.70**	**.22**	**.37**	**-.10**	**.10**	**-.11**	**.62**
32. is considerate and kind to almost everyone	.00	**.56**	.00	.00	.00	**.64**	.05	**.61**	**.12**	**.06**	.02	**.57**
37. is sometimes rude to others	.00	**.54**	.00	.00	.00	**.59**	**-.08**	**.47**	**.06**	**-.16**	**-.09**	**.60**
42. likes to cooperate with others	.00	**.38**	.00	.00	.00	**.78**	**.33**	**.34**	.01	**-.07**	.04	**.74**
3. does a thorough job*	.00	.00	**.76**	.00	.00	**.66**	.01	**.06**	**.76**	.03	-.01	**.64**
8. can be somewhat careless	.00	.00	**.41**	.00	.00	**.77**	-.04	-.06	**.42**	**-.19**	**-.09**	**.67**
13. is a reliable worker	.00	.00	**.74**	.00	.00	**.63**	.02	**.13**	**.72**	**.09**	.01	**.58**
18. tends to be disorganized*	.00	.00	**.58**	.00	.00	**.67**	-.01	**-.07**	**.59**	**-.08**	**-.09**	**.59**
23. tends to be lazy	.00	.00	**.40**	.00	.00	**.82**	**.14**	.00	**.36**	**-.16**	-.02	**.80**
28. perseveres until the task is finished*	.00	.00	**.64**	.00	.00	**.84**	.01	.04	**.62**	.00	**.10**	**.79**
33. does things efficiently*	.00	.00	**.74**	.00	.00	**.52**	.03	**.13**	**.70**	.03	**.10**	**.58**
38. makes plans and follows through with them	.00	.00	**.64**	.00	.00	**.33**	**.09**	.00	**.59**	-.04	**.14**	**.37**
43. is easily distracted	.00	.00	**.58**	.00	.00	**.51**	-.04	-.04	**.53**	**-.23**	-.03	**.51**
4. is depressed, blue*	.00	.00	.00	**.51**	.00	**.90**	**-.24**	-.06	-.03	**.48**	**.09**	**.79**
9. is relaxed, handles stress well	.00	.00	.00	**.72**	.00	**.59**	**-.07**	**-.10**	**-.05**	**.61**	-.04	**.58**
14. can be tense	.00	.00	.00	**.55**	.00	**.81**	.00	**-.09**	**.10**	**.62**	.03	**.74**
19. worries a lot	.00	.00	.00	**.48**	.00	**.82**	-.02	**.14**	**.10**	**.57**	.00	**.74**
24. is emotionally stable, not easily upset*	.00	.00	.00	**.69**	.00	**.81**	.02	**-.07**	**-.10**	**.61**	**-.06**	**.72**
29. can be moody	.00	.00	.00	**.43**	.00	**.69**	.05	**-.13**	**-.06**	**.47**	**.07**	**.59**
34. remains calm in tense situations	.00	.00	.00	**.62**	.00	**.46**	.02	-.02	**-.17**	**.51**	**-.15**	**.43**
39. gets nervous easily*	.00	.00	.00	**.51**	.00	**.62**	.01	**.14**	**.09**	**.59**	-.04	**.66**
5. is original, comes up with new ideas*	.00	.00	.00	.00	**.70**	**.95**	**.20**	-.04	**.06**	**-.10**	**.61**	**.94**
10. is curious about many different things	.00	.00	.00	.00	**.56**	**.43**	**.14**	.04	**.08**	-.04	**.48**	**.32**
15. is ingenious, a deep thinker*	.00	.00	.00	.00	**.60**	**.71**	**-.06**	-.02	**.09**	.02	**.65**	**.73**
20. has an active imagination	.00	.00	.00	.00	**.61**	**.60**	**.09**	**.06**	-.04	**.09**	**.62**	**.58**
25. is inventive	.00	.00	.00	.00	**.82**	**.74**	**.09**	**-.07**	.03	**-.12**	**.76**	**.65**
30. values artistic, aesthetic experiences	.00	.00	.00	.00	**.43**	**.77**	.03	**.16**	.01	**.15**	**.44**	**.70**
35. prefers work that is routine*	.00	.00	.00	.00	**.22**	**.84**	.00	**-.09**	-.05	**-.09**	**.23**	**.82**
40. likes to reflect, play with ideas	.00	.00	.00	.00	**.48**	**.85**	**-.11**	.05	.03	.05	**.55**	**.74**
41. has few artistic interests*	.00	.00	.00	.00	**.40**	**.67**	.00	**.07**	**-.08**	.00	**.42**	**.64**
44. is sophisticated in art, music, or literature	.00	.00	.00	.00	**.45**	**.79**	.00	**.10**	**-.07**	**.06**	**.49**	**.75**
Correlation with A	**.18**						**.08**					
Correlation with C	**.28**	**.23**					**.10**	**.11**				
Correlation with N	**-.25**	**-.23**	**-.26**				**-.09**	**-.04**	**-.15**			
Correlation with O	**.49**	.05	**.25**	**-.16**			**.21**	**.06**	**.14**	-.02		

Note. CFA = confirmatory factor analysis; ESEM = exploratory structural equation modeling; E = Extraversion, A = Agreeableness; C = Conscientiousness, N = Neuroticism, O = Openness; Uniq = Uniqueness. Items with an * are reverse-coded items. Bolded coefficients are statistically different from zero (p < .05); underlined coefficients in the ESEM solution are target loadings. Standard errors are available upon request from the corresponding author

Table 2 *Patterns of Mean Gender Differences on Big Five Observed and Latent Mean Factors and Correlations Among Observed, CFA and ESEM Big Five Inventory Scale Scores*

	E	*A*	*C*	*N*	*O*
	Observed scores				
Women (*M*±*SD*)	3.34±0.76	3.73±0.60	3.62±0.72	3.27±0.76	3.67±0.67
α	.83 (.81-.84)	.70 (.67-.73)	.82 (.81-.84)	.79 (.77-.81)	.82 (.80-.84)
Men (*M*±*SD*)	3.26±0.75	3.62±0.58	3.54±0.77	2.89±0.74	3.71±0.69
α	.80 (.78-.83)	.67 (.63-.71)	.85 (.83-.87)	.78 (.75-.81)	.80 (.78-.83)
Correlation with CFA scores (total sample)	.934 (.927-.940)	.979 (.977-.981)	.944 (.938-.949)	.972 (.969-.975)	.904 (.894-.913)
Correlation with ESEM scores (total sample)	.954 (.949-.959)	.960 (.956-.964)	.943 (.937-.949)	.968 (.965-.971)	.921 (.913-.929)
	ESEM Latent scores				
Correlation with CFA scores (total sample)	.948 (.942-.953)	.962 (.958-.966)	.990 (.989-.991)	.950 (.945-.955)	.978 (.976-.980)
MG5: FL + Int IN—Strong factorial/measurement IN	**-.17**	**-.24**	**-.13**	**-.62**	.05
MG5A: MG5 with CUs	**-.20**	**-.23**	**-.15**	**-.64**	.11
MG5B: MG5 with CUs IN	**-.20**	**-.24**	**-.15**	**-.64**	.10
MG7: FL + Int + Uniq IN—Strict factorial/measurement IN	**-.17**	**-.24**	**-.13**	**-.62**	.04
MG7A: MG7 with CUs	**-.20**	**-.23**	**-.15**	**-.64**	.10
MG7B: MG7 with CUs IN	**-.20**	**-.23**	**-.15**	**-.64**	.11
MG8: FL + FVCV + Int IN	**-.17**	**-.23**	**-.15**	**-.60**	.05
MG8A: MG8 with CUs	**-.20**	**-.23**	**-.17**	**-.63**	.11
MG8B: MG8 with CUs IN	**-.20**	**-.23**	**-.17**	**-.63**	.11
MG9: FL + FVCV +Int + Uniq IN	**-.17**	**-.24**	**-.15**	**-.60**	.04
MG9A: MG9 with CUs	**-.20**	**-.23**	**-.17**	**-.63**	.11
MG9B: MG9 with CUs IN	**-.20**	**-.23**	**-.17**	**-.63**	.11

Note. Women $n = 856$; Men $n = 530$; E = Extraversion, A = Agreeableness; C = Conscientiousness, N = Neuroticism, O = Openness. *M* = mean; SD = standard deviation*;* α = Cronbach's Alpha; bracketed figures show the 95% confidence interval.

See Table 3for a description of the models. Each of the 12 models provides estimates of standardized mean gender differences in the Big Five factors under different assumptions. MG = multiple group; FL = factor loadings; Inter = item intercepts; CUs = correlated uniquenesses; Uniq = item uniquenesses (error variances); IN = invariance; Bolded coefficients are statistically significant at $p < .05$; Negative coefficients indicate higher scores in females. Standard errors are are available upon request from the corresponding author

Table 3 *Summary of Goodness of Fit Statistics for All Gender Invariance (IN) Models*

Model and description	χ^2	df	CFI	TLI	NParm	RMSEA
MG1 – No invariance (Configural Invariance)						
MG1	4362.346	1472	.829	.780	596	.053
MG1A: MG1 with CUs	2647.178	1346	.923	.892	722	.037
MG1B: MG1 with CUs IN	2741.562	1409	.921	.894	659	.037
MG2: FL –Weak factorial/measurement IN (Nested with 1)						
MG2	4599.810	1667	.826	.803	401	.050
MG2A: MG2 with CUs	2910.672	1541	.919	.900	527	.036
MG2B: MG2 with CUs IN	3001.316	1604	.917	.902	464	.035
MG3: FL and Uniq (Nested with 1, 2)						
MG3	4701.686	1711	.823	.804	357	.050
MG3A: MG3 with CUs	3028.159	1585	.914	.898	483	.036
MG3B: MG3 with CUs IN	3107.559	1648	.913	.901	420	.036
MG4: FL + FVFC (Nested with 1, 2)						
MG4	4611.289	1682	.826	.805	386	.050
MG4A: MG4 with CUs	2922.527	1556	.919	.902	512	.036
MG4B: MG4 with CUs IN	3012.526	1619	.917	.903	449	.035
MG5: FL + Int – Strong factorial/measurement invariance (Nested with 1, 2)						
MG5	4743.794	1706	.820	.800	362	.051
MG5A: MG5 with CUs	3021.504	1580	.915	.898	488	.036
MG5B: MG5 with CUs IN	3110.629	1643	.913	.900	425	.036
MG6: FL +FVCV + Uniq (Nested with 1-4)						
MG6	4714.181	1726	.823	.806	342	.050
MG6A: MG6 with CUs	3039.967	1600	.915	.899	468	.036
MG6B: MG6 with CUs IN	3119.484	1663	.914	.902	405	.036
MG7: FL + Int + Uniq – strict factorial/measurement invariance (Nested with 1-3, 5)						
MG7	4848.671	1750	.816	.801	318	.051
MG7A: MG7 with CUs	3140.830	1624	.910	.895	444	.037
MG7B: MG7 with CUs IN	3218.307	1687	.909	.898	381	.036
MG8: FL + FVCV + Int (Nested with 1, 2 ,4, 5)						
MG8	4755.614	1721	.820	.802	347	.050
MG8A: MG8 with CUs	3033.525	1595	.915	.899	473	.036
MG8B: MG8 with CUs IN	3121.986	1658	.913	.901	410	.036
MG9: FL + FVCV +Int + Uniq (Nested with 1-8)						
MG9	4861.410	1765	.816	.803	303	.050
MG9A: MG9 with CUs	3152.796	1639	.910	.896	429	.037
MG9B: MG9 with CUs IN	3230.356	1702	.909	.899	366	.036
MG10: FL + Int + LFMn – latent mean IN (Nested with 1, 2, 5)						
MG10	4865.843	1711	.813	.793	357	.052
MG10A: MG10 with CUs	3154.608	1585	.907	.889	483	.038
MG10B: MG10 with CUs IN	3245.150	1648	.905	.891	420	.037
MG11: FL + Int + LFMn + Uniq – manifest mean IN (Nested with 1-3, 5, 7, 10)						
MG11	4969.482	1755	.809	.795	313	.051
MG11A: MG11 with CUs	3274.665	1629	.902	.887	439	.038
MG11B: MG11 with CUs IN	3352.215	1692	.902	.890	376	.038
MG12: FL + FVCV + Int + LFMn (Nested with 1, 2, 4-6, 8, 10)						
MG12	4877.117	1726	.813	.795	342	.051
MG12A: MG12 with CUs	3166.197	1600	.907	.890	468	.038
MG12B: MG12 with CUs IN	3255.923	1663	.906	.893	405	.037
MG13: FL + FVCV + Int + LFMn + Uniq – complete factorial IN (Nested with 1-12)						
MG13	4982.190	1770	.810	.796	298	.051
MG13A: MG13 with CUs	3286.739	1644	.903	.888	424	.038
MG13B: MG13 with CUs IN	3364.282	1707	.902	.891	361	.037

Note. Women n = 856; Men n = 530; χ^2 = model chi-square statistic; df = model degrees of freedom; CFI = comparative fit index; TLI = Tucker–Lewis index; NFParm = number of free parameters; RMSEA = root-mean-square error of approximation; MG (as in MG1) = multiple group; CUs = a priori correlated uniquenesses based on previous works; IN = the sets of parameters constrained to be invariant across the multiple groups, for MG invariance models; FL = factor loadings; Uniq = item uniquenesses (error variance); FVCV = factor variances–covariances; Int = item intercepts; LFMn = factor means.

Supplementary material for the paper *Testing the factor structure and measurement invariance across gender of the Big Five Inventory through Exploratory Structural Equation Modeling*

SM 01

Methodological Issues of Measurement Invariance

Measurement invariance across gender was tested through a 13-model taxonomy of invariance tests that integrates factor and measurement invariance traditions (Marsh et al., 2009). Following Meredith (1993), the sequence of invariance testing begins with a model of ‘configural’ invariance: that is, with no invariance of parameter estimates. That is, all parameters are freely estimated, such that only similarity in the overall pattern of parameters is evaluated. Since this model does not require any estimated parameters to be the same, it is not an actual invariance model, but its fit must be evaluated in order to provide both a test of the ability of the a priori model to fit the data in each group without invariance constraints, and a baseline for comparing other models that impose equality constraints on the parameter estimates across groups. The next step in invariance testing is to test a ‘weak’ measurement invariance model. This requires that factor loadings be invariant over groups. In fact, Byrne, Shavelson, and Muthén (1989) also suggest testing *partial invariance models* where, based on post-hoc modification indexes, some parameter estimates (e.g., factor loadings) are not constrained to be invariant.

If indicator means (i.e., the intercepts of responses to individual items) are also constrained to be equal across groups, then a ‘strong’ measurement invariance model is specified. If such model fits, factor loadings and item intercepts are invariant over groups, then changes in the latent factor means can reasonably be interpreted as changes in the latent constructs, since they are corrected for measurement error. The invariance of item intercepts is a critical issue, since it is an implicit assumption in the comparison of latent and manifest group means, but it has often been ignored and left untested in Big Five research (for a review and a discussion, see e.g., Marsh et al., 2010).

A finding in support of the invariance of item intercepts would entail that gender differences in latent scores based on each of the items considered separately were reasonably consistent in terms of magnitude as well as direction. A lack of invariance in item intercepts would mean that the latent group differences were not consistent across the items used to represent a latent factor on a particular scale (the so-called 'differential item functioning'), and would provide no basis for the generalizability of the results across a wider and more diverse set of items representing the trait (Marsh, Nagengast & Morin, 2012). Supplementary Material SM02 demonstrates that the sample at hand afforded sufficient statistical power (i.e., .80) to test this model. Besides, factor loadings and item intercept invariance are necessary but not sufficient conditions for testing *manifest* group mean differences, which also require invariance of item uniquenesses. The presence of differences in reliability (as represented or absorbed in the item uniquenesses) across groups could in fact distort mean differences on the observed scores. A model that specifies the invariance of item uniquenesses is referred to as a ‘strict’ invariance model.

The invariance of the factor variance-covariance matrix is not a prior focus of measurement invariance, but it is often crucial in studies of the invariance of covariance structures. Specifically, it is an important focus in studies that investigate the discriminant validity of multidimensional constructs that might subsequently be extended to include relations with other constructs. Typically, the comparison of correlations among FFA factors across groups is based on manifest scores that do not control for measurement error and that make implicit invariance assumptions that are rarely tested.

Recently, Marsh et al. (2009) expanded this measurement invariance tradition, suggesting a taxonomy of 13 partially nested models. Models vary from the least restrictive model of configural invariance to a model of complete invariance that posits strict invariance, together with invariance of the latent means and of the factor variance-covariance matrix (see Table 1 in the text and Marsh et al., 2009 for a more extended discussion of these issues).

Essentially, all models except configural invariance (Model 1) assume the invariance of factor loadings, but the invariance of indicator uniquenesses, for example, can be tested with or without the invariance of item intercepts. However, it must be noted that models with freely estimated indicator intercepts and freely estimated latent means are not identified. Hence, when intercepts are freely estimated, the latent means are fixed to be zero. In models that allow the estimation of differences in latent means, as explained by Sörbom (1974), it is not possible to estimate the latent means in both groups. Hence, the latent means are constrained to be zero in one group and are freely estimated in the second group: this means that the freely estimated latent mean, and its statistical significance, reflects the differences between the two groups.

Models of measurement invariance typically are tested within a CFA framework. In this study we used tests of measurement invariance over gender on the basis of a taxonomy of invariance tests within an ESEM framework.

References

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*, 456–466.

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.

Marsh, H. W., Muthén, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J. S., & Trautwein, U. (2009). Exploratory structural equation modeling, integrating CFA and EFA: Application to students' evaluations of university teaching. *Structural Equation Modeling, 16*, 439–476.

Marsh, H. W., Nagengast, B., & Morin, A. J. (2012). Measurement invariance of big-five factors over the life span: ESEM tests of gender, age, plasticity, maturity, and la dolce vita effects. *Developmental Psychology, 49*(6), 1194–1218.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance, *Psychometrika, 58*, 525–543.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology, 27*, 229–239.

SM02

Simulation Study for the Multiple Group ESEM with Measurement Invariance of Intercepts and Factor Loadings

Although the samples used in this study can be considered adequately large (856 female and 530 male participants), we tested whether or not it could be considered adequate for testing the measurement invariance of intercepts and factor loadings using the procedure described in Muthén and Muthén (2002). The method relies on Monte Carlo simulations in which data are generated from a population with hypothesized parameter values. Ten thousand samples are drawn, and a model is estimated for each sample. Parameter values and standard errors are averaged over the samples and the following criteria are examined: parameter estimate bias, standard error bias, and coverage. In this case we followed the guidelines provided by the *Mplus User's Guide* (Muthén & Muthén, 1998–2010), Example 12.12, with the following settings for starting values:

- .80 for target loadings in both groups
- .00 for cross-loadings in both groups
- 3.00 for intercepts in both groups
- 1.00 for factor variances in one group and 1.50 in the other
- .20 for factor correlations in one group and .50 in the other
- .60 for uniquenesses in one group and .80 in the other
- .00 for factor means in one group and .20 in the other

Muthén and Muthén (2002) suggest considering, as a first criterion, that parameter and standard error biases do not exceed 10% for any parameter in the model. The second criterion is that the standard error bias for the parameter for which power is being assessed does not exceed 5% (in this case we focused on factor means). The third criterion is that coverage remains between .91 and .98. Once these three conditions are satisfied, the sample size is chosen to keep power close to 0.80, a commonly accepted value for sufficient power. In our research the highest parameter bias was 6.54%, the highest standard error bias being 2.30% (2.0% for factor means); the coverage varied between .934 and .961. Hence, we can conclude that the sample we used afforded sufficient statistical power.

Muthén, B. & Muthén, L. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling, 4*, 599–620.

Muthén, L. K., & Muthén, B. (1998–2010). *Mplus user's guide*. Los Angeles, CA: Muthén & Muthén.

SM03

Standard Errors for Factor Loadings and Factor Correlations in Table 1

	CFA solution					ESEM solution				
Item	E	A	C	N	O	E	A	C	N	O
1. is talkative	.033	-	-	-	-	.025	.025	.021	.021	.023
6. is reserved*	.047	-	-	-	-	.031	.031	.027	.028	.028
11. is full of energy	.035	-	-	-	-	.032	.034	.029	.029	.031
16. generates a lot of enthusiasm	.026	-	-	-	-	.030	.034	.027	.027	.032
21. tends to be quiet	.035	-	-	-	-	.026	.024	.021	.023	.023
26. has an assertive personality*	.035	-	-	-	-	.032	.031	.026	.027	.028
31. is sometimes shy, inhibited	.037	-	-	-	-	.041	.033	.026	.031	.030
36. is outgoing, sociable*	.032	-	-	-	-	.020	.023	.018	.018	.021
2. tends to find fault with others*	-	.086	-	-	-	.028	.041	.026	.030	.029
7. is helpful and unselfish with others	-	.097	-	-	-	.030	.042	.026	.027	.029
12. starts quarrels with others	-	.092	-	-	-	.029	.047	.028	.031	.033
17. has a forgiving nature*	-	.037	-	-	-	.029	.032	.026	.028	.029
22. is generally trusting	-	.044	-	-	-	.031	.038	.028	.031	.029
27. can be cold and aloof	-	.065	-	-	-	.033	.047	.029	.035	.033
32. is considerate and kind to almost everyone	-	.072	-	-	-	.030	.041	.027	.028	.025
37. is sometimes rude to others	-	.073	-	-	-	.032	.046	.027	.033	.031
42. likes to cooperate with others	-	.070	-	-	-	.033	.033	.028	.029	.029
3. does a thorough job*	-	-	.018	-	-	.019	.020	.018	.020	.020
8. can be somewhat careless	-	-	.029	-	-	.029	.035	.031	.032	.030
13. is a reliable worker	-	-	.021	-	-	.022	.027	.023	.023	.023
18. tends to be disorganized*	-	-	.026	-	-	.026	.029	.026	.027	.025
23. tends to be lazy	-	-	.030	-	-	.035	.036	.035	.033	.031
28. perseveres until the task is finished*	-	-	.026	-	-	.025	.027	.027	.024	.025
33. does things efficiently*	-	-	.021	-	-	.022	.027	.022	.022	.023
38. makes plans and follows through with them	-	-	.024	-	-	.027	.029	.026	.025	.027
43. is easily distracted	-	-	.031	-	-	.027	.034	.034	.031	.029
4. is depressed, blue*	-	-	-	.027	-	.029	.030	.027	.027	.029
9. is relaxed, handles stress well	-	-	-	.026	-	.026	.027	.024	.032	.025
14. can be tense	-	-	-	.027	-	.025	.032	.026	.025	.025
19. worries a lot	-	-	-	.032	-	.025	.032	.024	.027	.025
24. is emotionally stable, not easily upset*	-	-	-	.023	-	.027	.028	.024	.031	.025
29. can be moody	-	-	-	.032	-	.029	.038	.028	.030	.030
34. remains calm in tense situations	-	-	-	.028	-	.027	.031	.025	.033	.029
39. gets nervous easily*	-	-	-	.031	-	.023	.033	.025	.027	.023
5. is original, comes up with new ideas*	-	-	-	-	.021	.028	.029	.024	.024	.026
10. is curious about many different things	-	-	-	-	.029	.030	.032	.028	.030	.033
15. is ingenious, a deep thinker*	-	-	-	-	.030	.025	.026	.022	.024	.030
20. has an active imagination	-	-	-	-	.024	.025	.025	.023	.024	.025
25. is inventive	-	-	-	-	.021	.025	.023	.021	.020	.029
30. values artistic, aesthetic experiences	-	-	-	-	.031	.028	.033	.027	.028	.031
35. prefers work that is routine*	-	-	-	-	.036	.032	.034	.030	.032	.040
40. likes to reflect, play with ideas	-	-	-	-	.035	.027	.032	.027	.027	.036
41. has few artistic interests*	-	-	-	-	.029	.028	.031	.028	.029	.032
44. is sophisticated in art, music, or literature	-	-	-	-	.027	.026	.029	.025	.027	.028
Correlation with A	.106					.022				
Correlation with C	.045	.044				.020	.019			
Correlation with N	.042	.073	.037			.022	.021	.019		
Correlation with O	.042	.065	.034	.035		.020	.024	.020	.020	

Note: E = Extraversion, A = Agreeableness; C = Conscientiousness, N = Neuroticism, O = Openness

SM04

Standard Errors for Latent Mean Differences Coefficients in Table 2

	E	*A*	*C*	*N*	*O*
MG5: FL + Int IN - Strong factorial/measurement IN	.063	.073	.058	.070	.065
MG5A: MG5 with CUs	.064	.074	.057	.072	.065
MG5B: MG5 with CUs IN	.064	.073	.058	.073	.064
MG7: FL + Int + Uniq IN - Strict factorial/measurement IN	.062	.071	.058	.069	.064
MG7A: MG7 with CUs	.062	.071	.057	.072	.065
MG7B: MG7 with CUs IN	.063	.071	.057	.072	.065
MG8: FL + FVCV + Int IN	.063	.073	.063	.065	.064
MG8A: MG8 with CUs	.065	.073	.063	.068	.063
MG8B: MG8 with CUs IN	.065	.073	.063	.068	.063
MG9: FL + FVCV +Int + Uniq IN	.063	.073	.063	.065	.064
MG9A: MG9 with CUs	.065	.073	.063	.068	.063
MG9B: MG9 with CUs IN	.065	.073	.063	.068	.064

Note: E = Extraversion, A = Agreeableness; C = Conscientiousness, N = Neuroticism, O = Openness; MG = multiple group; FL = factor loadings; Inter = item intercepts; CUs = correlated uniquenesses; Uniq = item uniquenesses (error variances); IN = invariance

SM05

Exploratory Structural Equation Modeling Standardized Factor Loadings, Uniquenesses, Intercepts and Factor Correlations Based on Responses to the Big Five Inventory in Females (n = 856) and Males (n = 530) from the Configural Invariance Model (see MG1B in Table 3)

	Females							Males						
Item	E	A	C	N	O	Uniq	Int	E	A	C	N	O	Uniq	Int
1. is talkative	**.67**	-.04	.01	**-.07**	**.07**	**.53**	**3.30**	**.66**	**.08**	-.01	**-.09**	.00	**.55**	**2.94**
6. is reserved*	**.51**	-.07	**-.19**	**-.12**	**-.11**	**.72**	**2.06**	**.39**	-.01	**-.25**	-.03	-.08	**.81**	**1.95**
11. is full of energy	**.41**	.06	**.21**	**.20**	**.16**	**.63**	**3.58**	**.36**	-.01	**.29**	**.18**	**.17**	**.64**	**3.53**
16. generates a lot of enthusiasm	**.52**	.05	.04	**.10**	**.26**	**.57**	**3.45**	**.49**	**.17**	.07	.02	**.16**	**.63**	**3.64**
21. tends to be quiet	**.61**	-.02	-.02	-.02	-.03	**.64**	**2.57**	**.70**	-.02	-.05	.01	**-.09**	**.54**	**2.37**
26. has an assertive personality*	**.52**	-.05	**.22**	**.16**	**.25**	**.47**	**3.53**	**.45**	-.04	**.23**	**.17**	**.22**	**.56**	**3.51**
31. is sometimes shy, inhibited	**.47**	**-.15**	-.04	**.21**	-.06	**.72**	**2.10**	**.40**	-.06	.06	**.28**	-.04	**.74**	**2.14**
36. is outgoing, sociable*	**.81**	**.11**	**.05**	.01	.01	**.30**	**3.51**	**.81**	.05	**-.08**	-.01	**.08**	**.31**	**3.06**
2. tends to find fault with others*	**-.11**	**.44**	-.04	**.26**	.02	**.73**	**2.63**	**-.16**	**.35**	.01	**.21**	**-.14**	**.79**	**2.73**
7. is helpful and unselfish with others	**.13**	**.47**	**.17**	**-.17**	.08	**.68**	**5.10**	**.19**	**.48**	**.08**	**-.12**	**.17**	**.65**	**4.00**
12. starts quarrels with others	**-.19**	**.48**	.02	**.15**	**-.11**	**.71**	**3.96**	**-.15**	**.32**	-.02	**.13**	-.05	**.86**	**4.07**
17. has a forgiving nature*	.01	**.50**	**-.14**	.01	.06	**.74**	**2.93**	-.07	**.53**	**-.12**	.03	.01	**.72**	**3.02**
22. is generally trusting	.02	**.44**	-.02	-.07	-.04	**.81**	**3.52**	-.05	**.45**	-.04	**.10**	-.07	**.79**	**3.50**
27. can be cold and aloof	**.27**	**.40**	**-.08**	-.06	**-.08**	**.77**	**2.62**	**.14**	**.31**	**-.17**	-.09	-.10	**.85**	**2.33**
32. is considerate and kind to almost everyone	.06	**.58**	**.12**	-.06	.00	**.63**	**4.38**	.06	**.62**	**.13**	**-.09**	.03	**.57**	**3.80**
37. is sometimes rude to others	**-.13**	**.51**	.03	**.15**	-.04	**.69**	**2.95**	-.03	**.41**	.07	**.20**	**-.14**	**.76**	**2.83**
42. likes to cooperate with others	**.37**	**.31**	-.01	**.08**	.04	**.74**	**3.82**	**.28**	**.40**	.04	.05	.02	**.72**	**3.62**
3. does a thorough job*	-.01	**.06**	**.74**	.00	.03	**.44**	**4.51**	.03	.03	**.79**	-.01	-.05	**.39**	**3.83**
8. can be somewhat careless	-.07	-.07	**.41**	**.19**	**-.10**	**.78**	**2.15**	-.02	-.02	**.42**	**.19**	-.08	**.76**	**2.18**
13. is a reliable worker	.04	**.14**	**.73**	**-.08**	.01	**.42**	**4.97**	.00	**.09**	**.73**	**-.14**	-.02	**.47**	**4.82**
18. tends to be disorganized*	-.01	-.07	**.57**	**.11**	**-.13**	**.65**	**2.84**	-.03	-.07	**.60**	**.10**	-.01	**.62**	**2.60**
23. tends to be lazy	**.15**	.07	**.33**	**.13**	-.03	**.82**	**2.22**	**.12**	-.09	**.39**	**.20**	.03	**.75**	**2.31**
28. perseveres until the task is finished*	.06	.07	**.59**	-.03	**.09**	**.60**	**3.59**	-.05	-.01	**.65**	.07	**.14**	**.52**	**3.41**
33. does things efficiently*	.05	**.12**	**.70**	-.02	**.10**	**.44**	**4.86**	.02	**.15**	**.71**	-.03	**.10**	**.43**	**4.38**
38. makes plans and follows through with them	**.12**	-.03	**.59**	.03	**.12**	**.59**	**3.84**	.03	.05	**.58**	**.11**	**.16**	**.56**	**3.54**
43. is easily distracted	**-.10**	-.06	**.53**	**.26**	-.04	**.62**	**2.40**	.03	-.01	**.51**	**.21**	-.01	**.66**	**2.34**
4. is depressed, blue*	**.23**	**.10**	.01	**.46**	**-.12**	**.69**	**3.01**	**.26**	-.02	.05	**.52**	-.02	**.63**	**3.38**
9. is relaxed, handles stress well	.06	**.12**	**.07**	**.61**	.03	**.57**	**2.25**	**.10**	.10	.05	**.55**	.03	**.64**	**2.69**
14. can be tense	.01	**.09**	**-.11**	**.61**	-.01	**.63**	**2.22**	-.02	**.15**	**-.09**	**.58**	-.07	**.64**	**2.40**
19. worries a lot	.04	**-.13**	**-.10**	**.58**	-.02	**.66**	**1.96**	-.04	**-.13**	**-.13**	**.55**	.03	**.71**	**2.11**
24. is emotionally stable, not easily upset*	-.01	**.08**	**.11**	**.61**	.04	**.57**	**2.40**	-.03	.08	**.09**	**.56**	.06	**.65**	**2.74**
29. can be moody	-.07	**.19**	**.08**	**.42**	-.02	**.76**	**1.89**	.01	.11	.07	**.44**	**-.15**	**.76**	**2.20**
34. remains calm in tense situations	.02	.02	**.16**	**.49**	**.15**	**.67**	**2.55**	-.05	.06	**.22**	**.45**	**.11**	**.69**	**3.09**
39. gets nervous easily*	-.01	**-.16**	-.05	**.56**	.03	**.68**	**2.17**	-.01	-.10	**-.14**	**.61**	.05	**.64**	**2.51**
5. is original, comes up with new ideas*	**.24**	-.07	.03	**.11**	**.61**	**.48**	**3.49**	**.17**	.03	**.11**	.05	**.58**	**.54**	**3.57**
10. is curious about many different things	**.14**	.06	**.15**	.05	**.48**	**.66**	**4.13**	**.12**	-.02	-.02	.06	**.53**	**.68**	**3.77**
15. is ingenious, a deep thinker*	-.03	-.03	**.13**	-.03	**.67**	**.52**	**3.40**	-.08	-.01	.05	-.05	**.64**	**.59**	**3.50**
20. has an active imagination	**.09**	**.07**	-.02	**-.12**	**.60**	**.60**	**4.04**	**.12**	.05	-.04	-.07	**.62**	**.57**	**4.03**
25. is inventive	**.09**	-.06	.00	**.12**	**.76**	**.37**	**3.46**	**.11**	**-.07**	.06	**.11**	**.79**	**.30**	**3.59**
30. values artistic, aesthetic experiences	.01	**.14**	.03	**-.15**	**.44**	**.76**	**3.88**	.03	**.19**	-.04	**-.11**	**.46**	**.73**	**3.36**
35. prefers work that is routine*	-.01	-.07	**-.08**	**.09**	**.27**	**.92**	**2.13**	.03	-.08	.01	.03	**.20**	**.95**	**2.23**
40. likes to reflect, play with ideas	**-.14**	.04	.04	-.05	**.55**	**.70**	**4.00**	-.06	.08	.03	-.06	**.55**	**.69**	**3.87**
41. has few artistic interests*	-.01	.05	**-.08**	.01	**.46**	**.79**	**2.92**	-.01	.10	**-.10**	.02	**.38**	**.84**	**2.70**
44. is sophisticated in art, music, or literature	-.02	.06	-.06	-.04	**.52**	**.73**	**2.77**	.04	**.18**	-.07	-.08	**.42**	**.77**	**2.67**
Correlation with A	**.06**							**.10**						
Correlation with C	**.11**	**.13**						**.07**	.07					
Correlation with N	**-.11**	**-.07**	**-.15**					**-.11**	-.06	**-.19**				
Correlation with O	**.20**	.05	**.14**	-.03				**.22**	.07	**.14**	.00			

Note. E = Extraversion, A = Agreeableness; C = Conscientiousness, N = Neuroticism, O = Openness; Uniq = Uniqueness; Int = Intercept. Items with an * are reverse-coded items. Bolded coefficients are statistically different from zero ($p < .05$); underlined coefficients in the ESEM solution are target loadings.

Exploratory Structural Equation Modeling Standard Errors for Standardized Factor Loadings, Uniquenesses, Intercepts and Factor Correlations Based on Responses to the Big Five Inventory in Females (n = 856) and Males (n = 530) from the Configural Invariance Model (see MG1B in Table 3)

	Females							Males						
Item	E	A	C	N	O	Uniq	Int	E	A	C	N	O	Uniq	Int
1. is talkative	.029	.028	.030	.027	.033	.039	.092	.032	.037	.036	.035	.038	.043	.099
6. is reserved*	.033	.035	.037	.038	.039	.037	.040	.045	.052	.045	.052	.053	.040	.048
11. is full of energy	.039	.037	.040	.039	.044	.032	.102	.048	.050	.051	.052	.058	.037	.113
16. generates a lot of enthusiasm	.040	.035	.037	.038	.045	.032	.095	.043	.050	.044	.047	.054	.040	.120
21. tends to be quiet	.031	.031	.031	.031	.033	.036	.060	.031	.035	.037	.033	.040	.043	.068
26. has an assertive personality*	.035	.034	.037	.034	.039	.033	.100	.042	.046	.050	.052	.055	.041	.113
31. is sometimes shy, inhibited	.038	.037	.047	.041	.046	.044	.040	.041	.045	.049	.056	.051	.044	.052
36. is outgoing, sociable*	.024	.024	.023	.023	.032	.032	.100	.026	.034	.031	.036	.033	.048	.107
2. tends to find fault with others*	.035	.034	.036	.043	.055	.043	.054	.044	.047	.052	.049	.057	.044	.073
7. is helpful and unselfish with others	.039	.039	.040	.041	.062	.049	.168	.039	.047	.038	.044	.059	.053	.175
12. starts quarrels with others	.038	.033	.036	.042	.057	.047	.147	.051	.054	.057	.052	.065	.043	.183
17. has a forgiving nature*	.037	.035	.037	.041	.041	.041	.079	.043	.042	.042	.047	.054	.057	.111
22. is generally trusting	.039	.037	.041	.041	.046	.039	.104	.042	.042	.049	.048	.056	.051	.128
27. can be cold and aloof	.040	.037	.043	.051	.055	.049	.063	.050	.054	.060	.056	.078	.051	.065
32. is considerate and kind to almost everyone	.034	.041	.041	.041	.057	.058	.130	.036	.038	.039	.039	.046	.054	.145
37. is sometimes rude to others	.038	.035	.040	.048	.065	.058	.067	.042	.048	.054	.050	.061	.049	.081
42. likes to cooperate with others	.038	.038	.042	.037	.042	.033	.108	.043	.043	.048	.050	.053	.043	.133
3. does a thorough job*	.026	.024	.025	.028	.027	.035	.152	.028	.034	.032	.030	.034	.040	.140
8. can be somewhat careless	.042	.043	.039	.046	.050	.036	.044	.045	.045	.048	.047	.052	.039	.057
13. is a reliable worker	.029	.029	.028	.031	.034	.037	.189	.034	.037	.036	.036	.043	.047	.204
18. tends to be disorganized*	.031	.035	.033	.037	.038	.038	.077	.040	.041	.044	.043	.051	.044	.085
23. tends to be lazy	.039	.046	.043	.046	.049	.031	.047	.049	.051	.052	.056	.057	.041	.063
28. perseveres until the task is finished*	.034	.036	.035	.033	.039	.040	.101	.038	.042	.040	.039	.042	.045	.122
33. does things efficiently*	.028	.028	.028	.029	.035	.035	.139	.035	.038	.037	.035	.043	.043	.164
38. makes plans and follows through with them	.032	.031	.034	.035	.039	.035	.109	.042	.047	.043	.045	.047	.043	.126
43. is easily distracted	.038	.046	.035	.046	.048	.046	.054	.042	.044	.046	.045	.051	.040	.065
4. is depressed, blue*	.036	.036	.038	.036	.038	.034	.033	.041	.046	.045	.048	.052	.042	.040
9. is relaxed, handles stress well	.032	.030	.033	.042	.030	.045	.062	.041	.044	.054	.044	.054	.054	.066
14. can be tense	.030	.033	.031	.034	.045	.038	.133	.044	.043	.045	.043	.049	.049	.133
19. worries a lot	.030	.029	.031	.033	.045	.037	.089	.041	.042	.039	.042	.050	.040	.096
24. is emotionally stable, not easily upset*	.031	.030	.034	.043	.034	.048	.056	.043	.041	.056	.046	.052	.060	.060
29. can be moody	.039	.038	.039	.045	.052	.037	.076	.045	.048	.048	.047	.058	.043	.058
34. remains calm in tense situations	.035	.033	.035	.045	.038	.041	.053	.043	.050	.054	.048	.057	.045	.062
39. gets nervous easily*	.029	.031	.030	.037	.042	.039	.075	.042	.042	.040	.040	.052	.047	.073
5. is original, comes up with new ideas*	.032	.032	.036	.032	.036	.031	.095	.037	.042	.041	.043	.046	.043	.124
10. is curious about many different things	.045	.038	.040	.040	.046	.040	.138	.044	.051	.051	.051	.052	.050	.149
15. is ingenious, a deep thinker*	.033	.029	.029	.029	.033	.042	.091	.037	.046	.044	.046	.045	.055	.128
20. has an active imagination	.032	.033	.030	.033	.033	.036	.126	.035	.047	.042	.047	.043	.051	.168
25. is inventive	.036	.027	.031	.028	.031	.049	.093	.035	.042	.031	.039	.036	.055	.136
30. values artistic, aesthetic experiences	.039	.033	.032	.036	.040	.034	.121	.046	.050	.048	.051	.057	.043	.118
35. prefers work that is routine*	.046	.038	.040	.042	.050	.026	.044	.052	.058	.054	.054	.052	.023	.064
40. likes to reflect, play with ideas	.043	.036	.034	.037	.040	.045	.123	.042	.055	.045	.045	.056	.057	.160
41. has few artistic interests*	.037	.035	.033	.038	.040	.033	.074	.047	.053	.051	.050	.055	.039	.090
44. is sophisticated in art, music, or literature	.030	.031	.030	.033	.034	.031	.069	.044	.048	.050	.050	.053	.039	.083
Correlation with A	.029							.035						
Correlation with C	.025	.023						.033	.042					
Correlation with N	.026	.026	.024					.033	.035	.031				
Correlation with O	.025	.030	.025	.025				.032	.041	.031	.031			

SM06

Models with Wording Method Factors

As an alternative approach to specifying correlated uniquenesses (CUs) among items belonging to the same Big Five facet, models with wording method factors—namely, straight- and reverse-coded items—were specified, in order to test whether they could also yield an adequate fit. The results suggest that the fit of these models was only marginally acceptable, and was worse than for models with CUs.

Model and description	χ^2	df	CFI	TLI	NFParm	RMSEA
Total group CFA with CUs (as reported in the paper)						
TGCFA1A: no CUs	5879.232	892	.700	.682	142	.064
TGCFA1B: CUs	3715.740	829	.826	.802	205	.050
Total group CFA with method factors						
Both method factors, uncorrelated	4036.984	848	.808	.786	186	.052
Both method factors, correlated	3983.629	847	.811	.789	187	.052
Straight items method factor	4569.577	864	.777	.756	170	.056
Reverse items method factor	5308.022	876	.773	.712	158	.060
Total group ESEM (as reported in the paper)						
TGESEM1A: no CUs	3415.029	736	.839	.793	298	.051
TGESEM1B: CUs	1823.265	673	.931	.903	361	.035
Total group ESEM with method factors						
Both method factors, uncorrelated	2251.764	342	.906	.872	342	.040
Both method factors, correlated	2261.251	343	.906	.871	343	.040
Straight items method factor	2762.233	708	.876	.835	326	.046
Reverse items method factor	2804.187	720	.875	.835	314	.046

Note. χ^2 = model chi-square statistic; df = model degrees of freedom; CFI = comparative fit index; TLI = Tucker–Lewis index; NFParm = number of free parameters; RMSEA = root-mean-square error of approximation; CFA = confirmatory factor analysis; CUs = a priori correlated uniquenesses based on previous works; ESEM = exploratory structural equation modelling.

SM07

Descriptive Statistics for the Big Five Inventory Items (n = 1,836)

Item	Min	Max	Mean	SD	SK	KU
bfi01	1	5	3.66	1.17	-0.61	-0.53
bfi02	1	5	3.15	1.18	0.03	-1.04
bfi03	1	5	4.07	0.97	-0.94	0.32
bfi04	1	5	2.19	1.22	0.63	-0.76
bfi05	1	5	3.59	1.02	-0.46	-0.31
bfi06	1	5	2.40	1.19	0.58	-0.57
bfi07	1	5	4.09	0.90	-1.01	1.02
bfi08	1	5	2.80	1.31	0.23	-1.13
bfi09	1	5	3.09	1.22	-0.03	-1.02
bfi10	1	5	4.06	1.02	-1.02	0.41
bfi11	1	5	3.70	1.05	-0.52	-0.39
bfi12	1	5	4.27	1.07	-1.38	0.99
bfi13	1	5	4.28	0.87	-1.24	1.32
bfi14	1	5	3.76	0.99	-0.77	0.29
bfi15	1	5	3.65	1.06	-0.51	-0.36
bfi16	1	5	3.50	1.00	-0.33	-0.18
bfi17	1	5	3.63	1.23	-0.61	-0.63
bfi18	1	5	3.58	1.31	-0.54	-0.90
bfi19	1	5	3.64	1.18	-0.56	-0.61
bfi20	1	5	3.98	0.99	-0.88	0.37
bfi21	1	5	3.26	1.31	-0.17	-1.10
bfi22	1	5	3.70	1.06	-0.66	-0.11
bfi23	1	5	3.02	1.34	0.04	-1.19
bfi24	1	5	2.93	1.23	0.05	-1.00
bfi25	1	5	3.64	1.04	-0.53	-0.27
bfi26	1	5	3.66	1.04	-0.51	-0.35
bfi27	1	5	3.29	1.33	-0.19	-1.16
bfi28	1	5	3.81	1.09	-0.65	-0.37
bfi29	1	5	3.32	1.36	-0.35	-1.12
bfi30	1	5	3.95	1.09	-0.85	-0.04
bfi31	1	5	2.54	1.20	0.56	-0.62
bfi32	1	5	3.91	0.95	-0.73	0.16
bfi33	1	5	3.96	0.85	-0.63	0.20
bfi34	1	5	2.81	1.18	0.24	-0.85
bfi35	1	5	2.98	1.38	0.05	-1.22
bfi36	1	5	3.75	1.13	-0.65	-0.41
bfi37	1	5	3.39	1.17	-0.18	-1.00
bfi38	1	5	3.75	1.01	-0.61	-0.19
bfi39	1	5	3.26	1.21	-0.26	-0.88
bfi40	1	5	3.97	1.01	-0.86	0.24
bfi41	1	5	3.58	1.27	-0.44	-0.91
bfi42	1	5	3.75	1.01	-0.59	-0.13
bfi43	1	5	3.01	1.27	-0.02	-1.08
bfi44	1	5	3.42	1.26	-0.35	-0.90
Min			2.19	0.85	-1.38	-1.22
Median			3.64	1.15	-0.52	-0.47
Max			4.28	1.38	0.63	1.32

Note: Min = Minimum; Max = Maximum; M = mean; SD = Standard Deviation; SK = skewness; KU = Kurtosis