Psychological Assessment
2010, Vol. 22, No. 3, 471–491

1040-3590/10/$12.00 DOI: 10.1037/a0019227

A New Look at the Big Five Factor Structure Through Exploratory Structural Equation Modeling

Herbert W. Marsh
University of Oxford

Oliver Lüdtke
University of Tübingen, Tübingen, Germany, and Max Planck Institute for Human Development, Berlin, Germany

Bengt Muthén
University of California, Los Angeles

Tihomir Asparouhov
Muthén & Muthén, Los Angeles, California

Alexandre J. S. Morin
University of Sherbrooke

Ulrich Trautwein
University of Tübingen, Tübingen, Germany, and Max Planck Institute for Human Development, Berlin, Germany

Benjamin Nagengast
University of Oxford

NEO instruments are widely used to assess Big Five personality factors, but confirmatory factor analyses (CFAs) conducted at the item level do not support their a priori structure due, in part, to the overly restrictive CFA assumptions. We demonstrate that exploratory structural equation modeling (ESEM), an integration of CFA and exploratory factor analysis (EFA), overcomes these problems with responses (N = 3,390) to the 60-item NEO–Five-Factor Inventory: (a) ESEM fits the data better and results in substantially more differentiated (less correlated) factors than does CFA; (b) tests of gender invariance with the 13-model ESEM taxonomy of full measurement invariance of factor loadings, factor variances–covariances, item uniquenesses, correlated uniquenesses, item intercepts, differential item functioning, and latent means show that women score higher on all NEO Big Five factors; (c) longitudinal analyses support measurement invariance over time and the maturity principle (decreases in Neuroticism and increases in Agreeableness, Openness, and Conscientiousness). Using ESEM, we addressed substantively important questions with broad applicability to personality research that could not be appropriately addressed with the traditional approaches of either EFA or CFA.

Keywords: exploratory structural equation modeling, factorial and measurement invariance, Big Five personality structure, differential item functioning

Supplemental materials: http://dx.doi.org/10.1037/a0019227.supp

Arguably, the most important advance in personality psychology in the past half century has been the emerging consensus that individual differences in adults' personality characteristics can be organized in terms of five broad trait domains: Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness. These Big Five factors now serve as a common language in the field, facilitating communication and collaboration. Although there are several Big Five instruments (e.g., Benet-Martinez & John, 1998; Caprara & Perugini, 1994; Goldberg, 1990; Gosling, Rentfrow, & Swann, 2003; John & Srivastava, 1999; Paunonen, 2003; Paunonen & Ashton, 2001; Saucier, 1998), the family of NEO instruments—including the 60-item NEO–Five-Factor Inventory (NEO-FFI; Costa & McCrae, 1992; McCrae & Costa, 2004) considered here—appear to be the most widely used instruments and to have received the most attention over recent years (Boyle, 2008).

Factor analysis has been at the heart of these exciting breakthroughs. Exploratory factor analyses (EFAs) have consistently identified the Big Five factors, and an impressive body of empirical research supports their stability and predictive validity (see McCrae & Costa, 1997). However, confirmatory factor analyses

Herbert W. Marsh and Benjamin Nagengast, Department of Education, University of Oxford, Oxford, England; Oliver Lüdtke and Ulrich Trautwein, University of Tübingen, Tübingen, Germany, and Center for Educational Research, Max Planck Institute for Human Development, Berlin, Germany; Bengt Muthén, Graduate School of Education and Information Studies, University of California, Los Angeles; Tihomir Asparouhov, Muthén & Muthén, Los Angeles, California; Alexandre J. S. Morin, Department of Psychology, University of Sherbrooke, Sherbrooke, Quebec, Canada.

This research was supported in part by a grant to Herbert W. Marsh from the United Kingdom's Economic and Social Research Council.

Correspondence concerning this article should be addressed to Herbert W. Marsh, Department of Education, University of Oxford, 15 Norham Gardens, Oxford OX2 6PY, United Kingdom. E-mail: herb.marsh@education.ox.ac.uk

(CFAs) have failed to provide clear support for the five-factor model on the basis of standard measures such as the NEO instruments. For example, in a particularly relevant study comparing EFA and CFA factor structures based on NEO–Personality Inventory (NEO-PI) responses, Vassend and Skrondal (1997) reported highly discrepant findings, leading them to conclude

> (i) that the original NEO-PI model as well as later EFA-based revisions are false or at least unsatisfactory, and (ii) that at present we do not know how the NEO-PI scales should be modeled with the aim of obtaining a common, acceptable NEO-PI version. (p. 157)

Problematic results based on CFAs have led some researchers to question the appropriateness of CFA for Big Five research (see Borkenau & Ostendorf, 1990; Church & Burke, 1994; McCrae, Zonderman, Costa, Bond, & Paunonen, 1996; Parker, Bagby, & Summerfeldt, 1993; Vassend & Skrondal, 1997). However, many of the methodological and statistical advances in quantitative psychology in the last 2 decades are associated with latent-variable approaches such as CFA and structural equation models (SEMs). Hence, failure to embrace these new and evolving methodologies (throwing the baby out with the bathwater) would have dire consequences—particularly for a field of research so fundamentally based on factor analysis. Indeed, assumptions of factorial and measurement invariance (in relation to multiple groups, time, covariates, and outcomes) that underpin nearly all Big Five studies cannot be appropriately evaluated with traditional approaches to EFA and thus have been largely ignored in Big Five EFA research. Here we outline a new approach to factor analysis—an integration of EFA and CFA—that has the potential to resolve this dilemma and has wide applicability to all disciplines of psychology that are based on the measurement of latent constructs. Thus, our study is a substantive-methodological synergy (Marsh & Hau, 2007), demonstrating the importance of applying new and evolving methodological approaches to substantively important issues.

Methodological Focus: Exploratory Structural Equation Modeling (ESEM)

EFA Versus CFA

Many measurement instruments used in psychological assessment apparently have well-defined EFA structures but are not supported by CFAs (Marsh et al., 2009). This concern led McCrae et al. (1996) to conclude:

> In actual analyses of personality data from Borkenau and Ostendorf (1990) to Holden and Fekken (1994), structures that are known to be reliable showed poor fits when evaluated by CFA techniques. We believe this points to serious problems with CFA itself when used to examine personality structure. (p. 568; also see Costa & McCrae, 1992, 1995; McCrae & Costa, 1997)

Church and Burke (1994) similarly concluded on the basis of their empirical research that

> Poor fits of a priori models highlighted not only the limited specificity of personality structure theory, but also the limitations of confirmatory factor analysis for testing personality structure models. (p. 93)

They argued that the independent clusters model (ICM) used in CFA studies, which requires each indicator to load on only one factor, is too restrictive for personality research, because indicators are likely to have secondary loadings unless researchers resort to using a small number of near-synonyms to infer each factor.

Marsh et al. (2009) claimed that, consistent with these concerns, many ad hoc strategies used to compensate for the inappropriateness of CFA in psychological research more generally are dubious, counterproductive, misleading, or simply wrong. Of particular relevance to the present investigation, the inappropriate imposition of zero factor loadings usually leads to distorted factors with positively biased factor correlations that might lead to biased estimates in SEMs incorporating other constructs (also see Marsh et al., 2009). In a similar vein, Marsh (2007; Marsh, Hau, & Grayson, 2005) concluded that many psychological instruments used in applied research do not even meet minimum criteria of acceptable fit according to current standards.

Apparently, many applied researchers persist with inappropriate ICM-CFA models because they believe that EFA approaches are outdated and that methodological advances associated with CFAs are not applicable to EFAs. Here we demonstrate how it is possible to apply EFA rigorously in a way that allows researchers to define more appropriately the underlying factor structure and to still apply the advanced statistical methods typically associated with CFAs and SEMs. This is accomplished with the ESEM procedure recently implemented in the Mplus statistical package (Version 5.2, Muthén & Muthén, 2008). Within the ESEM framework, the applied personality researcher has access to typical SEM parameter estimates, standard errors, goodness-of-fit statistics, and statistical advances normally associated with CFA and SEMs (see Asparouhov & Muthén, 2009; Marsh et al., 2009). Here we apply ESEM to NEO-FFI responses.

Tests of Factorial and Measurement Invariance

We know of no CFAs carried out at the item level—particularly for research based on the NEO-FFI instrument used to measure the Big Five personality factors—that provide acceptable support for the a priori Big Five factor structure. This is remarkable, given the widespread acceptance of the Big Five factor structure and the NEO-FFI. Hence, it is not surprising that research into the Big Five factor structure on responses to individual items continues to be based almost entirely on EFA (for exceptions, see Benet-Martinez & John, 1998; Dolan, Oort, Stoel, & Wichterts, 2009; Gustavsson, Eriksson, Hilding, Gunnarsson, & Ostensson, 2008; also see Reise, Smith, & Furr, 2001). We suggest that this failure to apply CFA models in Big Five research is due in large part to the inappropriateness of the typical ICM-CFA structure. Although identification of the appropriate factor structure is important in its own right, there are many other important advantages to the use of CFA that cannot be easily incorporated into EFA and thus have been largely ignored in Big Five personality research. Thus, for example, studies that use Big Five scale scores (or factor scores based on EFAs) are not corrected for measurement error. Although it is possible to correct for a simple form of measurement error (i.e., the typical correction for attenuation based on reliability estimates), in many applications the error structure is more complex (e.g., longitudinal studies as considered here), so the typical correction for attenuation is not sufficient.

A particularly important application of CFA techniques is to test the assumptions about the invariance of the Big Five factor structure over multiple groups or over time (Gustavsson et al., 2008; Nye, Roberts, Saucier, & Zhou, 2008; Reise et al., 2001). Unless the underlying factors are measuring the same construct in the same way and the measurements themselves are operating in the same way (across groups or over time), mean differences and other comparisons are likely to be invalid. Although some aspects of factor similarity can be addressed in part with EFA approaches (e.g., the similarity of the factor loadings), most cannot. In particular, an important assumption in the comparison of Big Five factors over different groups (e.g., men and women) or over time is the invariance of item intercepts. More specifically, it is important to ascertain that mean differences based on latent constructs (Big Five factors) are reflected in each of the individual items used to infer the latent constructs. For example, if the apparent level of gender differences in Extraversion varies substantially from item to item for different items used to infer this construct, then the gender differences based on the corresponding latent construct are idiosyncratic to the particular items used to infer Extraversion. Similarly, if responses to individual Extraversion items differ systematically with age (for different respondents) or over time (for the same respondents), then findings based on comparisons of scale scores might be invalid. In each case, these results would suggest that conclusions about differences in Extraversion do not generalize over even the set of items used in the instrument—let alone the population of items that could have been used. Hence, conclusions about differences in Extraversion might be idiosyncratic to the particular set of items and not be generalizable. From this perspective, it is important to evaluate the invariance of different aspects of the factor structure at the level of the individual item. Although issues of noninvariance of item intercepts (hereafter referred to as differential item functioning) are well known in evaluating the appropriateness of standardized achievement tests, these issues have been largely ignored in Big Five research (but see Jackson et al., 2009; Nye et al., 2008; Reise et al., 2001).

Substantive Focus on Big Five Personality Factors and the NEO-FFI

Gender Differences in Personality Traits

There is a long history of the search for gender differences in personality research (e.g., Feingold, 1994; Hall, 1984; Maccoby & Jacklin, 1974). Noting that Feingold (1994) had organized his review in part on the basis of the five broad factors and 30 facets of the NEO-PI, Costa, Terracciano, and McCrae (2001) greatly expanded the research based on the 30 facets measured by the NEO-PI-R for responses from 26 countries (N = 23,031). Interestingly, they found that gender differences within the set of six facets comprising each of the Big Five factors were not entirely consistent. Women had consistently higher scores across six facets representing Neuroticism and Agreeableness, whereas gender differences were consistently small for Conscientiousness. However, gender differences were less consistent for Extraversion and Openness; for each of these Big Five factors at least two (of six) facets favored women and at least two favored men. Hence, the size and even the direction of gender differences would differ depending on which facet (or mix of facets) was considered. Thus, even at the facet level there is apparently differential item (facet) functioning for some of the Big Five factors that compromises conclusions based on Big Five measures that are aggregated across facets. Logically, this implies that there is also likely to be differential item functioning at the level of individual items in relation to gender differences for NEO-FFI responses considered here.

Although there is considerable study-to-study variation in observed gender differences that may be a function of age, nationality, and the particular instrument considered, there is clear support for the conclusions that women tend to score higher than men in relation to Neuroticism and Agreeableness. Although less consistent, there is also evidence that women score higher on Conscientiousness and Extraversion but no clear support for evidence of gender differences in Openness. There is no evidence that men score higher than women on any of the Big Five factors as measured and labeled on the NEO-FFI (although women's higher scores on Neuroticism are sometimes summarized as lower scores on emotional stability). Particularly relevant to the current study (based on late-adolescent responses by Germans), Schmitt, Realo, Voracek, and Allik (2008) reported that for their German sample (N = 790), women scored higher than men did on all Big Five factors: Neuroticism (d = 0.48), Extraversion (d = 0.12), Agreeableness (d = 0.09), Conscientiousness (d = 0.23), and Openness (d = 0.11). Similarly, Donnellan and Lucas (2008) found that for the late-adolescent sample (ages 16–19 years) most relevant to the present investigation, German women consistently scored higher than German men did: Neuroticism (d = 0.47), Extraversion (d = 0.24), Agreeableness (d = 0.31), Conscientiousness (d = 0.34), and Openness (d = 0.36).

Longitudinal Invariance: Stability and Change in Personality Traits

The literature on personality development distinguishes several types of personality change and continuity (Caspi & Shiner, 2006; Lüdtke, Trautwein, & Husemann, 2009). Here we distinguish between correlational (rank-order), mean-level, and structural stability over time.

For correlational stability, cross-sectional and longitudinal research (Roberts & DelVecchio, 2000; see also Fraley & Roberts, 2005; Klimstra, Hale, Raaijmakers, Branje, & Meeus, 2009; Lüdtke et al., 2009) shows that correlational stability increases with age, particularly for the middle-to-late adolescent period that is the focus of the present investigation.

Studies of mean-level change with respect to life-span changes in Big Five traits show that most people become more dominant, agreeable, conscientious, and emotionally stable. Caspi, Roberts, and Shiner (2005) coined the term *maturity principle* to describe these findings of increasing psychological maturity from adolescence to middle age. In their meta-analysis of longitudinal studies, Roberts, Walton, and Viechtbauer (2006) also found substantial increases in Openness. For the 18–22 age group most relevant to the present investigation, Robins, Fraley, Roberts, and Trzesniewski (2001) found that, over a 4-year period, Agreeableness (d = 0.44), Conscientiousness (d = 0.27), and Openness (d = 0.22) increased and Neuroticism (d = –0.49) decreased. No statistically significant change was found for Extraversion. In summary, although results from these studies are not entirely consistent, there is general support for the maturity principle of increases in all Big Five factors (or decreases in Neuroticism instead of increases in Emotional Stability) except, perhaps, for Extraversion.

Structural stability assesses the extent to which the same factors are being assessed in different groups or over time. At least some level of structural invariance is a prerequisite for assessing either mean differences between groups or stability over time. If the nature of the factors changes so that factors are qualitatively different, then interpretations of stability over time are questionable. It is most appropriate to evaluate factorial and measurement invariance on the basis of responses to individual items. However, personality researchers have been remarkably unsuccessful in obtaining acceptable levels of goodness of fit for the a priori Big Five factor CFA structure when analysis of the structure is based on responses to individual items in studies of the NEO-FFI instrument considered here. Indeed, this might be considered the major limitation in Big Five personality research, particularly in relation to testing assumptions underpinning the valid assessment of stability over time as well as the valid comparison of latent means across groups. For this reason some studies have sought to formally test full measurement invariance based on mean responses averaged across different items, facet scores (e.g., Gignac, 2009; McCrae et al., 1996; Saucier, 1998; Small, Hertzog, Hultsch, & Dixon, 2003), parcel scores (Allemand, Zimprich, & Hendriks, 2008; Allemand, Zimprich, & Hertzog, 2007; Lüdtke et al., 2009; Marsh, Trautwein, Lüdtke, Köller, & Baumert, 2006), or scale scores (e.g., Mroczek & Spiro, 2003). Although these analyses are potentially useful, they have important limitations when conducted without prior verification of measurement invariance at the item level—an assumption underlying tests of mean differences (over time or across groups) and differential item functioning that could compromise the validity of interpretations based on analyses of aggregated scores (see later discussion for further elaboration). In the present investigation, we address these concerns, introducing a new ESEM approach that integrates the logic of the EFA approach typically used in Big Five personality research and the CFA approach widely argued to be inappropriate to Big Five research.

The Present Investigation: A Substantive-Methodological Synergy

Our study is a substantive-methodological synergy, demonstrating the power and flexibility of ESEM methods that integrate CFA and EFA (on the basis of the Mplus statistical package; Muthén & Muthén, 2008) to address substantively important issues about the Big Five factor structure on the basis of responses to the 60-item NEO-FFI instrument. We begin by comparing CFA and ESEM approaches, testing the assumption that ESEM models fit better than corresponding CFA models. For both CFA and ESEM models, we include 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 Big Five facet). Big Five theory posits that the Big Five factors should be reasonably orthogonal, but constraining all (nontarget) cross-loadings to be zero in the ICM-CFA model is posited to systematically inflate and bias estimates of the factor correlations. Hence, support for the prediction that Big Five factors are reasonably orthogonal is hypothesized to be stronger in ESEM models than in CFA models.

We then extend ESEM to test a 13-model taxonomy of measurement invariance, testing invariance of factor loadings, factor variances–covariances, item uniquenesses, CUs, item intercepts, and latent means—with a specific focus on gender differences in the latent means of the Big Five factors. Of particular interest are tests of the invariance of item intercepts that are an implicit assumption in the comparison of latent (or manifest) group means but are largely ignored in previous Big Five research (but see Jackson et al., 2009; Nye et al., 2008; Reise et al., 2001). We expect, on the basis of previous research, systematic differences, mostly reflecting higher means for women (particularly for the late-adolescent German sample considered here). We also predict that, consistent with previous research, there is differential item functioning in NEO-FFI responses (noninvariance of item intercepts) that would compromise the interpretation of latent mean comparisons, but we explore alternatives to circumvent this problem.

Finally, we apply ESEM to test–retest data, testing a set of models of measurement invariance over time with the inclusion of CUs relating responses to the same item on multiple occasions. Although these (within-group) tests of longitudinal invariance largely parallel those based on (between-group) tests over gender, the substantive implications are quite different. Indeed, given that participants are tested in their final year of high school at Time 1 (T1) and are tested 2 years after graduation at Time 2 (T2), it is reasonable that there might be systematic changes in Big Five latent means. We expect to see, based on the maturity principle, decreases in Neuroticism and increases in Agreeableness, Openness, and Conscientiousness.

Previous research has suggested a problem with the evaluation of stability over time for NEO-FFI responses that is especially relevant to the present investigation. NEO-FFI responses consistently have high levels of short-term test–retest stability (.86–.90; McCrae & Costa, 2004; Robins et al., 2001) and internal consistency (.68–.86; Costa & McCrae, 1992). However, this research suggests problems associated with a complex error structure in that test–retest correlations are larger than internal consistency measures of reliability. In particular, test–retest correlations would be greater than 1.0 if corrected for (internal consistency) unreliability. This suggests that observed test–retest correlations are more positively biased by CUs associated with specific variances of the same items administered on different occasions than negatively biased by the failure to control for measurement error in the factors. Traditional EFA approaches are unable to appropriately distinguish between measurement error on each occasion, CUs over time, and true stability of latent traits over time, but these issues can be addressed by ESEM, as demonstrated in the present investigation.

Method

Participants

The data come from a large, ongoing German study (Transformation of the Secondary School System and Academic Careers [TOSCA]; see Köller, Watermann, Trautwein, & Lüdtke, 2004; also see Lüdtke et al., 2009; Marsh, Trautwein, et al., 2006). A random sample of 149 upper secondary schools in a single German state was selected to be representative of the traditional and vocational gymnasium school types attended by the college-bound student population. At T1, the students (N = 3,390; 45% men, 55% women) were in their final year of upper secondary schooling (M age = 19.51, SD = 0.77). Two trained research assistants administered materials in each school, and students participated voluntarily, without any financial incentive. At T1, all students were asked to provide written consent to be contacted again later for a second wave of data collection. At T2, 2 years after graduation from high school, participants completed an extensive questionnaire taking about 2 hr in exchange for a financial reward of 10 euros (US$13).

For evaluation of longitudinal stability, our analyses are restricted to the responses by the 1,570 (39% men, 61% women) students who completed the NEO-FFI at both T1 and T2. To test for attrition effects, we compared continuers, who participated at both time points, to dropouts, who participated in only the first wave. Continuers had slightly lower grade point averages (M = 2.3 vs. 2.5) and were more likely to be female. Selectivity effects exceeding d = 0.10 were found for two of the Big Five scale scores; continuers had higher Conscientiousness and Agreeableness scores. Although dropouts and continuers differed statistically significantly in some domains, the magnitude of these differences was small and indicative of only small selectivity effects. We also compare, as part of the analysis, factor structures based on all students at T1 as well as those who completed instruments at both T1 and T2.

Measures: Big Five Dimensions

The 60-item NEO-FFI (Costa & McCrae, 1992) provides a short measure of the Big Five personality factors (Costa & McCrae, 1989). For each factor, McCrae and Costa (1989) selected 12 items from the 180 items of the longer NEO-PI (and the full 240-item NEO-PI-R), based primarily on correlations between each NEO-PI item and factor scores (McCrae & Costa, 1989). We measured the Big Five factors using the German version (Borkenau & Ostendorf, 1993) of the NEO-FFI, whose responses have high reliability, validity, and comparability with responses to the original English-language version (e.g., Borkenau & Ostendorf, 1993). In our study, items were rated on a 4-point scale ranging from 1 (*strongly disagree*) to 4 (*strongly agree*). Psychometric analyses of the 4-point response format show that this format has some advantages over a 5-point scale (Lüdtke, Trautwein, Nagy, & Köller, 2004). Coefficient alpha reliabilities at T1 and T2, respectively, were .78 and .80 (Extraversion), .72 and .73 (Agreeableness), .83 and .84 (Conscientiousness), .83 and .87 (Neuroticism), and .73 and .74 (Openness). Hence, consistent with previous research (e.g., Church & Burke, 1994; McCrae et al., 1996), there are small increases in reliability with increased age during this late-adolescent period.

Statistical Analyses

Analyses were conducted with Mplus (Version 5.2; Muthén & Muthén, 2008). Preliminary analyses consisted of a traditional CFA based on the Mplus robust maximum likelihood estimator (MLR), with standard errors and tests of fit that are robust in relation to nonnormality and nonindependence of observations (Muthén & Muthén, 2008). The main focus is on the application of ESEM to responses to the 60-item NEO Big Five personality instrument. The ESEM approach differs from the typical CFA approach in that all factor loadings are estimated, subject to constraints so that the model can be identified (for further details of the ESEM approach and identification issues, see technical appendix, Appendix 1 in the online supplemental materials; also see Asparouhov & Muthén, 2009). Here we used an oblique geomin rotation (the default in the Mplus) with an epsilon value of .5 and the MLR estimation. A critical advantage of the ESEM approach is the ability to test full measurement invariance for an EFA solution in relation to multiple groups or occasions.

Factorial and measurement invariance. Marsh et al. (2009) proposed a 13-model taxonomy of invariance tests that integrated factor analysis (e.g., Jöreskog & Sörbom, 1988; Marsh, 1994, 2007) and measurement invariance (e.g., Meredith, 1964, 1993; Meredith & Teresi, 2006) traditions to testing invariance over multiple groups or occasions. Following the measurement invariance tradition, we use terminology proposed by Meredith (1964, 1993) that has achieved broad acceptance. Although tests of invariance are frequently based on covariance matrices emerging from the factor analysis tradition, tests of full measurement invariance begin with raw data (or mean augmented covariance matrices) and should be done at the item level to evaluate item functioning.

In the Meredith (1964, 1993) tradition, the sequence of invariance testing generally begins with a model with no invariance of any parameter estimates (i.e., all parameters are freely estimated) such that only similarity of the overall pattern of parameters is evaluated (configural invariance). Technically, this model might not be an invariance model in that it does not require any estimated parameters to be the same. However, it does provide both a test of the ability of the a priori model to fit the data in each group (or occasion) without invariance constraints and a baseline for comparing other models that do impose equality constraints on the parameter estimates across groups or over time. Configural invariance models are followed by tests of *weak measurement invariance* that are satisfied if factor loadings are invariant over groups or occasions, although Byrne, Shavelson, and Muthén (1989) also argued for the usefulness of a less demanding test of partial invariance in which some parameter estimates are not constrained to be invariant. *Strong measurement invariance* is satisfied if the indicator means (i.e., the intercepts of responses to individual items) and factor loadings are invariant over groups. If 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. *Strict measurement invariance* is satisfied if factor loadings, item intercepts, and item uniquenesses are all invariant across groups or over time. Strict measurement invariance is required in order to compare Big Five (manifest) scale scores (or factor scores) over time or across different groups. As comparisons based on latent constructs are corrected for measurement error, they require only strong measurement invariance.

The taxonomy of 13 partially nested models (Marsh et al., 2009) expand this measurement invariance tradition; models vary from the least restrictive model of configural invariance with no invariance constraints to a model of complete invariance that posits strict invariance as well as the invariance of the latent means and of the factor variance–covariance matrix (see Table 1; for a more extended discussion of these issues, see also Marsh et al., 2009). All models except the configural invariance model (Model 1) assume

Table 1
Taxonomy of Invariance Tests for Evaluating Measurement Invariance of Big Five Responses Across Multiple Groups or Over Multiple Occasions

Model	Parameters constrained to be invariant
1	None (configural invariance)
2	FL [1] (weak factorial/measurement invariance)
3	FL, Uniq [1, 2]
4	FL, FVCV [1, 2]
5	FL, Inter [1, 2] (strong factorial/measurement invariance)
6	FL, Uniq, FVCV [1–4]
7	FL, Uniq, Inter [1–3, 5] (strict factorial/measurement invariance)
8	FL, FVCV, Inter [1, 2, 4, 5]
9	FL, Uniq, FVCV, Inter [1–8]
10	FL, Inter, LFMn [1, 2, 5] (latent mean invariance)
11	FL, Uniq, Inter, LFMn [1–3, 5, 7, 10] (manifest mean invariance)
12	FL, FVCV, Inter, LFMn [1, 2, 4–6, 8, 10]
13	FL, Uniq, FVCV, Inter, LFMn [1–12] (complete factorial invariance)

Note. Models with freely estimated LFMn constrain intercepts to be invariant across groups, whereas models in which intercepts are free imply that mean differences are a function of intercept differences. Values in brackets represent nesting relations in which the estimated parameters of the less general model are a subset of the parameters estimated in the more general model under which it is nested. All models are nested under Model 1 (with no invariance constraints), whereas Model 13 (complete invariance) is nested under all other models. FL = factor loadings; Uniq = item uniquenesses; FVCV = factor variances–covariances; Inter = item intercepts; LFMn = latent factor means. Parts of this table were adapted from "Exploratory Structural Equation Modeling, Integrating CFA and EFA: Application to Students' Evaluations of University Teaching," by H. W. Marsh, B. Muthén, T. Asparouhov, O. Lüdtke, A. Robitzsch, A. J. S. Morin, and U. Trautwein, 2009, *Structural Equation Modeling, 16,* p. 443, Table 1. Copyright 2009 by Taylor & Francis.

the invariance of factor loadings, but it is possible to test, for example, the invariance of indicator uniquenesses with or without the invariance of item intercepts. However, models with freely estimated indicator intercepts and freely estimated latent means are not identified. So in models with freely estimated intercepts, the latent means are fixed to be zero. Then, when the intercepts are constrained to equality across groups (or occasions), the latent means are constrained to be zero in one group (or occasion) and freely estimated in the second group (or occasion). In this manner, the latent means in the second group (or occasion) and their statistical significance reflect the differences between the two groups (or occasions).

Here we demonstrate the application of tests of measurement invariance over gender and across time on the basis of our taxonomy of invariance tests (see Table 1). Such tests have typically used SEM/CFA. Related multiple-group methods have been proposed for EFA (e.g., Cliff, 1966; Meredith, 1964), but they mainly focus on the similarity of factor patterns rather than formal tests of invariance (but also see Dolan et al., 2009). However, the ESEM model can be extended to multiple groups or longitudinal analyses such that the ESEM solution is estimated separately for each group or occasion and parameters can be constrained to be invariant across groups or over time (Marsh et al., 2009; also see technical appendix, Appendix 1 in the supplemental materials).

CUs. In general, the use of ex post facto CUs should be avoided (e.g., Marsh, 2007), but there are some circumstances in which a priori CUs should be included. When the same items are used on multiple occasions, there are likely to be correlations between the unique components of the same item administered on the different occasions that cannot be explained in terms of correlations between the factors. Indeed, Marsh and Hau (1996; Marsh, 2007), Jöreskog (1979), and others have argued that the failure to include these CUs is likely to systematically bias parameter estimates such that test–retest correlations among matching latent factors are systematically inflated, which can then systematically bias other parameter estimates (especially in SEMs). In the extreme, test–retest correlations might be so substantially inflated that the failure to include appropriate CUs can result in improper solutions such as a nonpositive definite factor variance–covariance matrix or estimated test–retest correlations that are greater than 1.0 (e.g., Marsh, Martin, & Debus, 2001; Marsh, Martin, & Hau, 2006). Previous research showed that short-term test–retest correlations for NEO-FFI factors are systematically larger than internal consistency estimates of reliability so that disattenuated test–retest correlations would be greater than 1.0 (see earlier discussion). This suggests that there are likely to be substantial CUs test–retest data considered here. For this reason, Marsh and Hau argued that CUs relating responses to the same items on different occasions should always be included in the a priori model, but it is easy to evaluate the extent to which the exclusion of these a priori CUs affects the fit of the model and the nature of parameter estimates (particularly test–retest stability coefficients) by constraining them to be zero. Importantly, it is difficult to either test or correct complex structures of measurement error with EFAs and scale scores typically used in Big Five research.

As described in more detail by McCrae and Costa (2004), in the NEO-PI-R (with 240 items), each of the Big Five factors was represented by six facets, and each facet was represented by multiple items. However, in the construction of the (short) NEO-FFI, items were selected to best represent each of the Big Five factors without reference to the facets. More specifically, each Big Five factor was represented by a factor score (based on an EFA with varimax rotation), and items were selected that were most highly correlated with this factor score. Hence, some facets are overrepresented (relative to the design of the full NEO-PI-R), whereas other facets are represented by a single item or not represented at all. We posited that items that came from the same facet of a specific Big Five factor would have higher correlations than would items that came from different facets of the same Big Five factor—beyond correlations that could be explained in terms of the common Big Five factor that they represented. Here we modeled these potentially inflated correlations due to facets as CUs relating each pair of items from the same facet. Based on the mapping of NEO-FFI items onto the NEO-PI-R facets (R. McCrae, personal communication, December 1, 2008; also see Appendix 2 of the supplemental materials), this resulted in an a priori set of 57 CUs inherent to the design of the NEO-FFI. Although we argue that this set of a priori CUs should be included in all factor analyses of NEO-FFI responses, we systematically evaluate models with and without these CUs as well as the invariance of these CUs over multiple (gender) groups and over time.

Goodness of fit. CFA/SEM research typically focuses on the ability of a priori models to fit the data as summarized by sample size independent indices of fit (e.g., Marsh, 2007; Marsh, Balla, & Hau, 1996; Marsh, Balla, & McDonald, 1988; Marsh et al., 2005). Here we consider the root-mean-square error of approximation (RMSEA), the Tucker–Lewis index (TLI), and the comparative fit index (CFI), as operationalized in Mplus in association with the MLR estimator (Muthén & Muthén, 2008). We also considered the robust chi-square test statistic and evaluation of parameter estimates. For both the TLI and CFI, values greater than .90 and .95, respectively, typically reflect acceptable and excellent fit to the data. For the RMSEA, values less than .05 and .08 reflect a close fit and a reasonable fit to the data, respectively (Marsh, Hau, & Wen, 2004). However, we emphasize that these cutoff values constitute only rough guidelines; there is considerable evidence that realistically large factor structures (e.g., instruments with at least 50 items and at least five factors) are typically unable to satisfy even the minimally acceptable standards of fit (Marsh, 2007; Marsh et al., 2005; also see Marsh, Hau, Balla, & Grayson, 1998). However, because there are few applications of ESEM—and none that fully evaluate the appropriateness of the traditional CFA indices of fit—it is unclear how relevant these CFA indices and proposed cutoff values are for ESEM studies (Marsh et al., 2009).

In CFA studies it is typically more useful to compare the relative fit of a taxonomy of nested (or partially nested) models designed a priori to evaluate particular aspects of interest than to compare that of single models (Marsh, 2007; Marsh et al., 2009). Any two models are nested so long as the set of parameters estimated in the more restrictive model is a subset of the parameters estimated in the less restrictive model. This comparison can be based on a chi-square difference test, but this test suffers the same problems as the chi-square test used to test goodness of fit that led to the development of fit indices (see Marsh et al., 1998). For this reason, researchers have posited a variety of ad hoc guidelines to evaluate when differences in fit are sufficiently large to reject a more parsimonious model (i.e., the more highly constrained model with fewer estimated parameters) in favor of a more complex model. It has been suggested that support for the more parsimonious model requires a change in CFI of less than .01 (Chen, 2007; Cheung & Rensvold, 2001) or a change in RMSEA of less than .015 (Chen, 2007). Marsh (2007) noted that some indices (e.g., TLI and RMSEA) incorporate a penalty for parsimony so that the more parsimonious model can fit the data better than a less parsimonious model can (i.e., the gain in parsimony is greater than the loss in fit). Hence, a more conservative guideline is that the more parsimonious model is supported if the TLI or RMSEA is as good as or better than that for the more complex model. Nevertheless, all these proposals should be considered as rough guidelines or rules of thumb.

Especially in relation to the taxonomy of invariance tests, support for the invariance of a set of parameters should be based in part on the similarity of parameters in models that do not impose invariance constraints as well as on the goodness of fit in models that do. Here we focus on both the similarity of the patterns of parameters and the levels of the parameter estimates. For example, here we evaluate the similarity of factor loadings on the basis of various CFA and ESEM models—whether the same item has a relatively high or low factor loading across different groups (or occasions)—with a profile similarity index (PSI). To compute the PSI, we simply construct a column that contains all the factor loadings for one group and a second column of corresponding factor loadings for the second group and then correlate the values from the two columns. Hence the PSI is merely the correlation between the two sets of factor loadings. To evaluate levels of the parameter estimates, we compare descriptive statistics for the set of coefficients in each group. Ultimately, however, an evaluation of goodness of fit must be based upon a subjective integration of many sources of information, including fit indices, a detailed evaluation of parameter estimates in relation to a priori hypotheses, previous research, and common sense.

Results

Big Five Factor Structure: ESEM Versus CFA

The starting point for the present investigation is to test our a prior hypothesis that the ESEM model provides a better fit to NEO-FFI responses than does a traditional ICM-CFA model. Indeed, as emphasized by Marsh et al. (2009), the ESEM analysis is predicated on the assumption that ESEM performs noticeably better than does the ICM-CFA model in terms of goodness of fit (see Table 2) and the construct validity of the interpretation of the factor structure.

The ICM-CFA solution does not provide an acceptable fit to the data (CFI = .685, TLI = .672; see TGCFA1A in Table 2), consistent with previous research. The next model (TGCFA1B) incorporates a priori CUs (based on the facet structure of the NEO-PI-R; see earlier discussion and Appendix 2 of the supplemental materials); results are still inadequate, albeit improved

Table 2
Summary of Goodness-of-Fit Statistics for Total Group Models (Time 1 Data)

Model and description	χ^2	*df*	CFI	TLI	NFParm	RMSEA
Total group CFA						
TGCFA1A: no CUs; no gender	15,488	1700	.685	.672	190	.049
TGCFA1B: CUs; no gender	12,567	1643	.750	.731	247	.044
Total group ESEM						
TGESEM1A: no CUs; no gender	8,013	1480	.851	.821	410	.036
TGESEM1B: CUs; no gender	5,201	1423	.914	.893	467	.028

Note. CFI = comparative fit index; TLI = Tucker–Lewis index; NFParm = number of free parameters; RMSEA = root-mean-square error of approximation; CFA = confirmatory factor analysis; ESEM = exploratory structural equation modeling; CUs = a priori correlated uniquenesses (based on the facet design of the instrument).

(CFI = .750, TLI = .731). The corresponding ESEM solutions fit the data much better. Although the fit of the total group with no a priori CUs is still not acceptable (TGESEM1A: CFI = .851, TLI = .821; see Table 2), the inclusion of CUs results in a marginally acceptable fit to the data (TGESEM1B: CFI = .914, TLI = .893, RMSEA = .028).

It is also instructive to compare parameter estimates based on the ICM-CFA and ESEM solutions (see Appendix 3 of the supplemental materials). In both types of models, the factor loadings tend to be modest, with few factor loadings greater than .70 and some factor loadings less than .30. Although CFA factor loadings (*Mdn* = .47) are slightly higher than those for the ESEM model (*Mdn* = .46), the differences are typically small and the pattern of factor loadings is similar for the CFA and ESEM solutions. To quantify this subjective evaluation, we computed a PSI in which the vector of 60 CFA factor loadings was related to the corresponding vector of 60 EFA target loadings. The PSI (r = .87) demonstrated that ESEM and CFA factor loadings were highly related. Consistent with McCrae and Costa (2004), the 14 items that they noted as potentially weak also had lower factor loadings than the remaining 56 items did for both ICM-CFA (M = .38 vs. .49, respectively) and ESEM (M = .32 vs. .48, respectively) solutions. Although a few of these 14 items performed well here, we note that these same items also did well in the original McCrae and Costa study. Importantly, almost all 60 items load more positively on the ESEM factor that each was designed to measure and less positively on all other factors.

A detailed evaluation of the factor correlations among the Big Five factors demonstrates a critical advantage of the ESEM approach over the ICM-CFA approach. Although patterns of correlations are similar, the CFA factor correlations (–.502 to +.400; *Mdn* absolute value = .197) tend to be systematically larger than the ESEM factor correlations (–.205 to +.140; *Mdn* absolute value = .064). Thus, for example, the negative correlation between Neuroticism and Extraversion is –.502 on the basis of the CFA solution but only –.205 for the ESEM solution. Similarly, the correlation between Extraversion and Conscientiousness is +.400 for the CFA results but only +.104 for the ESEM results. In this respect, the ESEM solution is more consistent with a priori predictions that the Big Five personality factors are reasonably orthogonal.

Clearly the ESEM solution is superior to the CFA solution, in terms of both fit and distinctiveness of the factors that are consistent with Big Five theory. The comparison of results from these two models provides the initial and most important test for the appropriateness of the ESEM model—at least relative to the CFA model. It is also important to emphasize that the goodness of fit for the ESEM model is apparently far better than what has ever been achieved in previous research with the NEO-FFI on the basis of factor analyses conducted at the item level.

Invariance Over Gender

How stable is the NEO-FFI factor structure over gender? Are there systematic gender differences in latent means, and are the underlying assumptions that are needed to justify interpretations of these results met? To address these questions, we applied our taxonomy of 13 ESEM models (see Table 1). The basic strategy is to apply the set of 13 models designed to test different levels of factorial and measurement invariance, ranging from the least demanding model, which imposes no invariance constraints (configural invariance), to the most demanding model, which posits complete gender invariance in relation to the Big Five factor structure, latent means, and item intercepts. However, application of this taxonomy of models is complicated by two features that are partially idiosyncratic to this application: the a priori CUs and tests of partial invariance of item intercepts (Byrne et al., 1989). The results already presented on the basis of the total sample indicate that a priori CUs are necessary to achieve even a minimally acceptable fit to the data. However, it is also important to determine the extent to which these a priori CUs are invariant over gender and how these influence the behavior of the various models.

For all 13 models we begin by evaluating the 57 a priori CUs. Hence, we first test models with no CUs (e.g., MG1 in Table 3 corresponds to the first model in the invariance taxonomy in Table 1). We then test two additional variations: one in which the a priori CUs are allowed to vary for men and women (submodels labeled *A* in the Description column of Table 3, as in MG1A) and another in which the CUs are constrained to be invariant over responses by men and women (submodels labeled *B* in Table 3, as in MG1B). Hence, within this set of three submodels there is a systematic nesting to evaluate the a priori CUs and their invariance over gender in relation to each of the 13 invariance models described in Table 1.

For the models that posit gender differences in latent means for the Big Five factors, we also test several models to evaluate partial invariance. Submodels labeled *C* posit partial invariance (i.e., item intercepts identified in preliminary analyses are freely estimated and not constrained to be invariant over gender—see subsequent discussion) but with no CUs. In submodels labeled *D* the set of 57 a priori CUs is added, and in submodels labeled *E* these a priori CUs are constrained to be equal over gender. Hence, within this set of five submodels there is a systematic nesting that allows evaluation of the CUs and their invariance over gender, partial invariance, and combinations of these constraints.

Model MG1 (see Table 3), with no invariance constraints, does not provide an acceptable fit to the data (TLI = .823, CFI = .852). Indeed, these fit statistics are approximately the same as those based on the total group ESEM model (see TGESEM in Table 2) with twice the degrees of freedom (2960 vs. 1480) and twice the number of estimated parameters (820 vs. 410). However, consistent with earlier results, the inclusion of the set of a priori CUs substantially improves the fit to a marginally acceptable level (TLI = .891, CFI = .912; see MG1A in Table 3). Importantly, constraining these a priori CUs to be invariant over gender (see MG1B in Table 3) resulted in almost no change in fit. For fit indices that control for parsimony, the fit is essentially unchanged or slightly better for MG1B than for MG1A, respectively (.891 to .892 for TLI; .028 to .028 for RMSEA). For the CFI that is monotonic with parsimony, the change (.912 to .911) is clearly less than the .01 value typically used to support invariance constraints. These results are substantively important, demonstrating that the sizes of the 57 a priori CUs are reasonably invariant over gender. For each of the 13 models used to test the factorial invariance of the full mean structure (see Table 1), the inclusion of this set of a priori CUs substantially improves the goodness of fit to a similar degree. Furthermore, for each of these tests comparing freely estimated CUs and constraining CUs to be invariant over gender, there is support for the invariance of the CUs. The consistency of

Table 3
Summary of Goodness-of-Fit Statistics for All Gender Invariance (IN) Models (Time 1 Data)

Model and description	χ^2	*df*	CFI	TLI	NFParm	RMSEA
MG1 (configural IN)						
MG1: no IN (configural IN)	9,373	2960	.852	.823	820	.036
MG1A: MG1 with CUs (not invariant over sex)	6,654	2846	.912	.891	934	.028
MG1B: MG1A with CUs IN (invariant over sex)	6,743	2903	.911	.892	877	.028
MG2 (FL; weak factorial/measurement IN)						
MG2: IN = FL (weak factorial/measurement IN)	9,831	3235	.848	.833	545	.035
MG2A: MG2 with CUs	7,124	3121	.908	.895	659	.028
MG2B: MG2A with CUs IN	7,218	3178	.907	.896	602	.027
MG3 (FL & Uniq)						
MG3: IN = FL, Uniq	10,264	3295	.839	.827	485	.035
MG3A: MG3 with CUs	7,513	3181	.900	.889	599	.028
MG3B: MG3A with CUs IN	7,644	3238	.898	.889	542	.028
MG4 (FL & FVCV)						
MG4: IN = FL, FVCV	9,908	3250	.846	.833	530	.035
MG4A: MG4 with CUs	7,204	3136	.906	.894	643	.028
MG4B: MG4A with CUs IN	7,296	3193	.905	.895	587	.028
MG5 (FL & Inter; strong factorial/measurement IN)						
MG5: IN = FL, Inter (strong factorial/measurement IN)	10,937	3290	.824	.810	490	.037
MG5A: MG5 with CUs	7,982	3176	.889	.876	604	.030
MG5B: MG5A with CUs IN	8,079	3233	.888	.878	547	.033
MG5C: MG5 with P-IN, no CUs	9,951	3267	.846	.833	513	.035
MG5D: MG5C with CUs	7,223	3153	.906	.895	627	.028
MG5E: MG5D with CUs IN	7,316	3210	.905	.895	570	.027
MG6 (FL, FVCV, Uniq)						
MG6: IN = FL, FVCV, Uniq	10,346	3310	.838	.826	470	.035
MG6A: MG6 with CUs	7,602	3196	.898	.887	584	.029
MG6B: MG6A with CUs IN	7,731	3253	.897	.888	527	.028
MG7 (FL, Uniq, Inter; strict factorial/measurement IN)						
MG7: IN = FL, Uniq, Inter (strict factorial/measurement IN)	11,377	3350	.815	.804	430	.038
MG7A: MG7 with CUs	8,376	3236	.881	.870	544	.031
MG7B: MG7A with CUs IN	8,505	3293	.880	.871	487	.031
MG7C: MG7 with Inter (P-IN), no CUs	10,383	3327	.837	.827	453	.035
MG7D: MG7C with CUs	7,611	3213	.899	.888	567	.028
MG7E: MG7D with CUs IN	7,744	3270	.897	.888	510	.028
MG8 (FL, FVCV, Inter)						
MG8: IN = FL, FVCV, Inter	11,012	3305	.822	.809	475	.037
MG8A: MG8 with CUs	8,060	3191	.888	.875	589	.030
MG8B: MG8A with CUs IN	8,156	3248	.887	.877	532	.030
MG8C: MG8 with Inter (P-IN), no CUs	10,029	3282	.844	.832	498	.035
MG8D: MG8C with CUs	7,303	3168	.905	.893	612	.028
MG8E: MG8D with CUs IN	7,397	3225	.904	.894	555	.028
MG9 (FL, Uniq, FVCV, Inter)						
MG9: IN = FL, FVCV, Uniq, Inter	11,458	3365	.813	.803	415	.038
MG9A: MG9 with CUs	8,464	3251	.880	.869	529	.031
MG9B: MG9A with CUs IN	8,591	3308	.878	.870	472	.031
MG9C: MG9 with Inter (P-IN), no CUs	10,467	3342	.836	.826	438	.035
MG9D: MG9C with CUs	7,700	3228	.897	.887	552	.029
MG9E: MG9D with CUs IN	7,829	3285	.895	.887	495	.029
MG10 (FL, Inter, LFMn; latent mean IN)						
MG10: IN = FL, Inter, LFMn	11,550	3295	.809	.795	485	.039
MG10A: MG10 with CUs	8,625	3181	.874	.860	599	.032
MG10B: MG10A with CUs IN	8,720	3238	.873	.862	542	.032
MG10C: MG10 with Inter (P-IN), no CUs	10,466	3272	.834	.820	508	.036
MG10D: MG10C with CUs	7,749	3158	.894	.881	622	.029
MG10E: MG10D with CUs IN	7,842	3215	.893	.882	565	.029
MG11 (FL, Uniq, Inter, LFMn; manifest mean IN)						
MG11: IN = FL, Uniq, Inter, LFMn	11,990	3355	.801	.790	425	.039
MG11A: MG10 with CUs	9,020	3241	.867	.854	539	.032
MG11B: MG10A with CUs IN	9,149	3298	.865	.855	482	.032
MG11C: MG10 with Inter (P-IN), no CUs	10,902	3332	.825	.814	448	.037
MG11D: MG10C with CUs	8,141	3218	.886	.875	562	.030
MG11E: MG10D with CUs IN	8,272	3275	.885	.875	505	.030

(table continues)

Table 3 (*continued*)

Model and description	χ^2	*df*	CFI	TLI	NFParm	RMSEA
MG12 (FL, FVCV, Inter, LFMn)						
MG12: IN = FL, FVCV, Inter, LFMn	11,638	3310	.808	.794	470	.039
MG12A: MG12 with CUs	8,717	3196	.873	.859	584	.032
MG12B: MG12A with CUs IN	8,812	3253	.872	.860	527	.032
MG12C: MG12 with Inter (P-IN), no CUs	10,552	3287	.832	.819	493	.036
MG12D: MG12C with CUs	7,838	3173	.892	.888	607	.029
MG12E: MG12D with CUs IN	7,931	3230	.892	.881	550	.029
MG13 (FL, Uniq, FVCV, Inter, LFMn; complete factorial IN)						
MG13: IN = FL, Inter, Uniq, FVCV, LFMn	12,084	3370	.799	.789	410	.039
MG13A: MG13 with CUs	9,121	3256	.865	.853	524	.033
MG13B: MG13A with CUs IN	9,249	3313	.863	.854	467	.033
MG13C: MG13 Inter (P-IN), no CUs	10,994	3347	.824	.813	433	.037
MG13D: MG13C with CUs	8,240	3233	.884	.873	547	.030
MG13E: MG13D with CUs IN	8,368	3290	.883	.873	490	.030

Note. For multiple-group (MG) IN models, IN refers to the sets of parameters constrained to be invariant across the multiple groups. CFI = comparative fit index; TLI = Tucker–Lewis index; NFParm = number of free parameters; RMSEA = root-mean-square error of approximation; CUs = correlated uniquenesses; FL = factor loadings; Uniq = item uniquenesses; FVCV = factor variances–covariances; Inter = item intercepts; P-IN = partial IN; LFMn = latent factor means.

this pattern of results over the wide variety of different models is impressive and provides clear support for the inclusion of these a priori CUs based on the design of the NEO-FFI. However, in order to facilitate communication of the results, we will focus primarily on models in which CUs are included and constrained to be invariant over gender (e.g., Model MG1B for Model 1).

Descriptive similarity of solutions for men and women. Before formally testing the invariance of different parameters over gender, it is useful to evaluate the similarity of solutions when these parameters are freely estimated for men and women (see Appendix 4 of the supplemental materials). Of particular importance are the factor loadings. First we evaluate how similar the pattern of factor loadings is for men and women based on a PSI (i.e., the relation between the 300 factor loadings based on responses by men and those based on responses by women). The extremely high PSI (r = .97) indicates that the pattern of factor loadings is similar. Furthermore, the actual values of the factor loadings are similar across the two groups. Nontarget loadings are consistently small for both groups (Men: –.33 to +.32, *Mdn* = –.01; Women: –.38 to +.32, *Mdn* = –.01), whereas target loadings were consistently higher (Men: .05 to .74, *Mdn* = .46; Women: .10 to .73, *Mdn* = .46). Although there are apparently a few weak items, even these items are typically weak across both groups. The pattern of factor correlations for the two groups is also similar (PSI = .93), whereas the absolute values of the correlations are consistently small (Men: .01 to .20, *Mdn* = .06; Women: .00 to .25, *Mdn* = .06). Item uniquenesses are also similar for the two groups (PSI = .91), as are the values for the two groups (Men: .43 to .99, *Mdn* = .72; Women: .47 to .99, *Mdn* = .73).

The invariance of item intercepts is especially important for subsequent tests of measurement invariance. The pattern of item intercepts is similar for the two groups (PSI = .94), but intercepts are somewhat higher for women (2.49 to 6.32, *Mdn* = 3.46) than for men (3.52 to 5.95, *Mdn* = 3.42). A nominal test of the significance of this difference was statistically significant (M for men = 3.52, M for women = 3.83), $t(59) = 7.15$, $p < .001$ (similar tests of significance on each of the other sets of parameters were nonsignificant). These differences in intercepts are consistent with higher mean ratings by women, but more appropriate tests of this observation require more formal tests of mean structure invariance pursued in the next section.

In summary, descriptive summaries of parameter estimates in Appendix 4 of the supplemental materials suggest that the factor solutions—with the possible exception of item intercepts—are similar for the two groups. We now pursue formal tests of this invariance in relation to the taxonomy of invariance models presented in Table 1.

Tests of invariance over gender. *Weak factorial/measurement invariance* tests whether the factor loadings are the same for men and women. Model MG2B (along with MG2 and MG2A) tests the invariance of factor loadings over gender. The critical comparison between the more parsimonious MG2B (with factor loadings invariant) and 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 = .896 vs. .892; RMSEA = .027 vs. .028), whereas the difference in CFI (.907 vs. .911) is less than the value of .01 typically used to reject the more parsimonious model.

Strong measurement invariance requires that item intercepts—as well as factor loadings—be invariant over groups. The critical comparison is thus between Models MG2B and MG5B and tests whether differences in the 60 intercepts can be explained in terms of five latent means (i.e., a complete absence of differential functioning). The change in df = 55 represents the 60 new constraints on item intercepts minus the five latent factor means that are now freely estimated. However, the fit of MG5B (CFI = .888, TLI = .878) is not acceptable and is worse than the fit of the corresponding model MG2B (CFI = .907, TLI = .896). Hence, gender differences at the level of item means cannot be explained in terms of the factor means, and there is differential item functioning between gender groups.

Because there is strong evidence that item intercepts are not completely invariant and invariance of item intercepts is so central to the evaluation of latent mean differences, we pursued alternative tests of partial invariance of item intercepts (see Models MG5C–

MG5E in Table 3). We identified, on the basis of (ex post facto) modification indices in which we freed parameters one at a time, 23 (of 60) item intercepts that contributed most to the misfit associated with the complete invariance of item intercepts (see Appendix 2 of the supplemental materials). The results support partial invariance of item intercepts. For example, fit indices that control for parsimony are nearly the same for MG5E compared with MG2B (.895 vs. .896 for TLI; .027 vs. .027 for RMSEA), whereas the difference in CFIs (.905 vs. .907) is less than the .01 value that would lead to the rejection of constraints imposed in MG5E. However, the interpretation of these results is cautioned by ex post facto modifications (see subsequent discussion about partial invariance).

Strict measurement invariance requires that item uniquenesses, item intercepts, and factors loadings all be invariant over the groups. Here, the critical comparison is between Models MG5 and MG7; the change in $df = 60$ represents the 60 new constraints for item uniquenesses. Although Model MG7B does not provide an adequate goodness of fit to the data, the addition of the ex post facto partial-invariance strategy for the intercepts substantially improves the fit. However, the fit of MG7E (CFI = .897, TLI = .888) is only marginally acceptable and is apparently worse than the fit of the corresponding model MG5E (CFI = .905, TLI = .895). However, comparison of all the various pairs of models that test this invariance of the uniquenesses (MG3B vs. MG2B; MG6B vs. MG4B; MG7B vs. MG5B; MG7E vs. MG5E; MG9B vs. MG8B; MG9E vs. MG8E; MG11B vs. MG10B; MG13B vs. MG12B; MG13E vs. MG12E) consistently results in a change in CFIs that is slightly less than the .01 value typically used to support the more parsimonious model with uniquenesses invariant. Although it would be possible to pursue a strategy of partial invariance of uniquenesses, we did not do so because the evaluation of latent mean differences that is our main focus does not depend on the invariance of uniquenesses.

Factor variance–covariance invariance is typically not a focus of measurement invariance, but it is frequently an important focus of studies of the invariance of covariance structures—particularly studies of the discriminant validity of multidimensional constructs that might subsequently be extended to include relations with other constructs. Although the comparison of correlations among Big Five factors across groups is common, these are typically based on manifest scores that do not control for measurement error and make implicit invariance assumptions that are rarely tested. Here, the most basic comparison is between Models MG2 (factor loadings invariant) and MG4 (factor loadings and factor variance–covariance invariant). The change in $df = 15$ represents the 10 factor covariances and five factor variances. The results provide reasonable support for the additional invariance constraints, both in terms of the values for the fit indices and their comparison with MG2. For example, fit indices that control for parsimony are nearly the same for MG4B compared with MG2B (.895 vs. .896 for TLI; .028 vs. .027 for RMSEA), whereas the difference in CFIs (.905 vs. .907) is less than the .01 cutoff value that would lead to the rejection of constraints imposed in MG4B.

Tests of the invariance of the latent factor variance–covariance matrix, as is the case with other comparisons, could be based on any pair of the six models in Table 3 that differ only in relation to whether the factor variance–covariance matrix is free or not. Although each of these pairs of models differs by $df = 15$, corresponding to the parameters in the variance–covariance matrix, they are not equivalent; support for the invariance of the variance–covariance matrix could be found in some of those comparisons but not in others. Although we suggest that the comparison between Models MG4 and MG2 is the most basic comparison, valuable information can also be obtained from the other comparisons as well. Especially if there are systematic, substantively important differences in the interpretations on the basis of these different comparisons, further scrutiny would be warranted in that true differences in the factor variance–covariance matrix might be "absorbed" into differences in other parameters that are not constrained to be invariant. Fortunately, this complication is not evident in the present investigation, because support for the invariance of factor variance–covariance matrix is consistent across each of these alternative comparisons.

Finally, we are now in a position to address the issue of the invariance of the factor means across the two groups. The final 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. Again, there are several models that could be used to test gender mean invariance; they include (a) MG5 versus MG10, (b) MG7 versus MG11, (c) MG8 versus MG12, and (d) MG9 versus MG13. However, our earlier inspection of item intercepts suggests that there are systematic gender differences in latent means. Hence, it is not surprising that Models 10–13 are also rejected. These results imply that latent means representing the Big Five factors differ systematically for men and women. Consistent with a priori predictions, latent means are systematically higher for women on all Big Five latent means, although the largest differences are for Neuroticism and Conscientiousness.

An alternative, pragmatic approach to the comparison of the means for the different models is to evaluate the extent to which the pattern of latent mean gender differences vary as a function of the models considered. Hence, in Table 4 we summarize gender differences on the basis of each of the 24 models that provide estimates of gender differences. The set of 276 PSIs among all possible pairs of the 24 profiles varied from .852 to .999 (mean r = .957). Therefore, the pattern of gender differences was similar across the different models. This suggests, at least in this application, that gender differences are reasonably robust in relation to violations of underlying assumptions of gender invariance in the various models.

Invariance Over Time

With some adaptation, it is possible to apply the same set of 13 models to test the invariance of the Big Five factor structure over time using the ESEM approach with test–retest data. As with the tests of invariance over gender, we hypothesized that the same set of 57 a priori CUs (based on the design of the NEO instrument) are required. Because there are parallel CUs for T1 and T2 responses, we can also test the invariance of these CUs over time. However, we also posit a second a priori set of 60 CUs to account for the residual associations between matching items at T1 and T2 (see earlier discussion). Here we distinguish within-wave CUs (WWCUs) and cross-wave CUs (CWCUs). The WWCUs consist of 57 WWCUs that are specific to the design of the NEO-FFI already considered in previous analyses. In the longitudinal models

Table 4
Patterns of Gender Differences on Big Five Latent Mean Factors

Model and description	NEUR	EXTR	OPEN	AGRE	CONC
MG5 (strong factorial/measurement IN)					
MG5: IN = FL, Inter	.622	.317	.378	.173	.597
MG5A: MG5 with CUs	.647	.330	.363	.156	.660
MG5B: MG5A with CUs IN	.646	.330	.361	.157	.660
MG5C: MG5 with P-IN, no CUs	.524	.436	.362	.289	.571
MG5D: MG5C with CUs	.553	.429	.333	.306	.598
MG5E: MG5D with CUs IN	.552	.430	.334	.307	.596
MG7 (strict factorial/measurement IN)					
MG7: IN = FL, Uniq, Inter	.621	.322	.381	.176	.600
MG7A: MG7 with CUs	.642	.338	.365	.159	.667
MG7B: MG7A with CUs IN	.643	.337	.364	.158	.667
MG7C: MG7 with P-IN, no CUs	.525	.443	.365	.294	.576
MG7D: MG7C with CUs	.551	.439	.335	.312	.605
MG7E: MG7D with CUs IN	.551	.437	.335	.311	.603
MG8					
MG8: IN = FL, FVCV, Inter	.680	.285	.374	.163	.579
MG8A: MG8 with CUs	.706	.294	.361	.156	.641
MG8B: MG8A with CUs IN	.708	.292	.358	.156	.641
MG8C: MG8 with P-IN, no CUs	.586	.405	.359	.281	.552
MG8D: MG8C with CUs	.614	.398	.332	.302	.577
MG8E: MG8D with CUs IN	.614	.398	.332	.302	.576
MG9					
MG9: IN = FL, FVCV, Uniq, Inter	.680	.287	.374	.164	.577
MG9A: MG9 with CUs	.706	.297	.358	.156	.639
MG9B: MG9A with CUs IN	.707	.295	.357	.155	.641
MG9C: MG9 with P-IN, no CUs	.588	.408	.359	.283	.553
MG9D: MG9C with CUs	.615	.401	.331	.305	.578
MG9E: MG9D with CUs IN	.614	.400	.330	.304	.578

Note. See Tables 1 and 2 for a description of the models. Each of the 28 models provides estimates of gender differences in the Big Five factors under different assumptions. The pattern of gender differences across the 28 models is similar, with the correlation varying from .848 to .999 (mean r = .959). NEUR = Neuroticism; EXTR = Extraversion; OPEN = Openness; AGRE = Agreeableness; CONC = Conscientiousness; MG = multiple group; IN = invariance (for multiple-group IN models, IN refers to the sets of parameters constrained to be invariant across the MGs); FL = factor loadings; Inter = item intercepts; CUs = correlated uniquenesses; P-IN = partial IN; Uniq = item uniquenesses.

considered here, we also posit that the same set of WWCUs affect responses at T1 and T2, and we test their invariance over time. CWCUs are the set of 60 CWCUs relating uniquenesses associated with matching items at T1 and T2. In these longitudinal models, we evaluate the effect of their inclusion on goodness of fit and on other parameter estimates in the model—particularly latent test–retest correlations of the same construct over time.

Longitudinal factor structure of NEO-FFI responses. *Configural invariance* refers to tests of whether the a priori model fits the data when no invariance constraints are imposed (see LIM1 in Table 5). In LIM1, no CUs are posited (neither WWCUs nor CWCUs) and the fit of LIM1 is poor (CFI = .737, TLI = .712). In LIM1A, the inclusion of the 60 CWCUs improves the fit substantially (CFI = .886, TLI = .874,) but is still not acceptable. In LIM1B, the two sets of 57 WWCUs (but not CWCUs) are added to Model LIM1 and then constrained to be invariant over time in LIM1C. Based on goodness of fit, there is a modest increase in fit associated with the addition of WWCUs and little or no decrement in fit associated with holding them invariant over the two waves of data. However, both of these models are technically improper in that the factor variance–covariance matrix is not positive definite (suggesting that some single latent variable or combination of latent variables is a linear combination of some other variable or some different combination of variables). Clearly this dictates caution in the interpretation of the results or, perhaps, that this model should simply be rejected as misspecified. Although these problems support our contention that CWCUs should be included, we return to this issue shortly.

In Model LIM1D, all the a priori CUs are included (the two sets of WWCUs and the one set of CWCUs). Then, in LIM1E, the two sets of WWCUs are constrained to be invariant over time. Unlike in the previous two longitudinal models, solutions based on these models are fully proper, represent a substantial improvement in goodness of fit over previous models, and are at least marginally acceptable in terms of goodness of fit (TLIs and CFIs are greater than .90). Furthermore, Model LIM1E provides good support for the invariance of the WWCUs over time (T1 and T2).

It is also instructive to compare the parameter estimates based on T1 and T2 ESEM solutions (see Appendix 4 of the supplemental materials). The sizes of the factor loadings tend to be modest, with few factor loadings greater than .70 and some target factor loadings less than .30. However, the pattern of loadings is similar across the two waves (PSI = .98). Although T2 target loadings (.10 to .72, *Mdn* = .50) are slightly higher than the T1 target loadings (.05 to .72, *Mdn* = .48), the differences are small. For both waves of data, the average nontarget loading is close to zero but quite variable (T1: –.43 to .27, *Mdn* = .00; T2: –.41 to .26, *Mdn* = .00). Also, the pattern of correlations among the 10 T1 factor correlations is similar to the matching T2 factor correlations

(PSI = .954). In each case, the absolute value of correlations is modest (T1: *Mdn r* = .096; T2: *Mdn r* = .088). Finally, the pattern of intercepts is also similar (PSI = .966), although T1 intercepts are consistently somewhat lower than those at T2 (T1: *Mdn* = 3.56, *M* = 3.75; T2: *Mdn* = 3.61, *M* = 3.83). Particularly results for T1 responses are similar to those considered earlier (see Table 2), but this is hardly surprising, because the T1 responses considered here are a subset of the data considered earlier. What is important, however, is that the factor solution for T1 is highly similar to that based on T2 responses by the same students. Next we pursue more formal tests of these observations for ESEM models of longitudinal invariance. On the basis of our initial analyses, primarily submodel E, which includes CWCUs and invariant WWCUs, is considered.

Invariance of NEO-FFI factor structure over time. *Weak factorial/measurement invariance* tests the invariance of factor loadings over time. Because model LIM2E (with factor loadings invariant over time) is so much more parsimonious than is LIM1E (factor loadings free), it is not surprising that the CFI is marginally better for LIM1E (.912) than for LIM2E (.907; see Table 5). However, this difference is less than the .01 difference typically taken as support for the less parsimonious model. Furthermore, indices that take into account parsimony (TLI and RMSEA) are nearly identical for the two models. Consistent with this observation, factor loadings for T1 and T2 when invariance constraints were not imposed were very similar (see earlier discussion).

Strong measurement invariance requires that item intercepts—as well as factor loadings—be invariant over time, and the critical comparison is between Models LIM2E (factor loadings invariance) and LIM5E (factor loadings and item intercepts invariant). The CFI for LIM5E (.899) is marginally lower than those for LIM2E (ΔCFI = .008) and particularly LIM1E (ΔCFI = .013), and these differences approach or exceed the nominal .01 cutoff. This difference is also evident in differences in TLIs that control for parsimony (.893 vs. .901 and .902 for LIM5E, LIM2E, and LIM1E, respectively). These results indicate that there is only modest support for invariance of item intercepts and suggest that there might be differential item functioning over time. Furthermore, this pattern of results is replicated in the comparison of other models that differ only in terms of intercept invariance (e.g., LIM8E vs. LIM4E, LIM9E vs. LIM6E). Because the invariance of item intercepts is so central to the evaluation of latent mean differences, we pursued alternative tests of partial invariance of item intercepts. We identified, on the basis of (ex post facto) modification indices, 11 (of 60) item intercepts that contributed most to the misfit associated with the complete invariance of item intercepts. We conclude, on the basis of submodel LIM5Ep (the *p* indicating partial invariance; CFI = .904, TLI = .898), that there is at least reasonable support for the partial invariance of item intercepts. Although the improved fit of this submodel (LIM5Ep) over the corresponding submodel of full intercept invariance (LIM5E) is not large, for now we focus on models of partial intercept invariance (based on freeing these 11 item intercepts) rather than complete intercept invariance (but return to this issue in subsequent discussion).

Strict measurement invariance requires that item uniquenesses, as well as item intercepts and factor loadings, be invariant over time. The critical submodel LIM7Ep tests the invariance of factor loadings and item uniquenesses and partial invariance of item intercepts (CFI = .899, TLI = .894). Consistent with interpretations of previous models, comparison of this submodel LIM7Ep with model LIM5Ep suggests modest support for the invariance of item uniquenesses (ΔCFI = .005, ΔTLI = .004). Additional comparisons of models differing only by the inclusion of invariant items' uniquenesses support this conclusion. Although it would be possible to pursue tests of partial invariance of uniquenesses, we did not do so as the evaluation of latent mean differences does not depend on the invariance of uniquenesses.

Tests of the invariance of the latent factor variance–covariance matrix, as is the case with other comparisons, could be based on any pair of models in Table 5 that differ only in relation to whether the factor variance–covariance matrix is free or not. The most basic comparison (LIM4E vs. LIM2E) suggests good support for the invariance of the factor variance–covariance matrix (ΔCFI = .000, ΔTLI = .000). Other pairs of models in Table 5 that differ only in relation to whether the factor variance–covariance matrix is free or not also show good support for the invariance of the factor variance–covariance matrix over time (also see related test–retest correlations in Table 6).

Finally, we are now in a position to address the issue of the invariance of the latent factor means over time. Submodels LIM10Ep–LIM13Ep each test the invariance of latent mean differences in combination with the invariance of other parameter estimates. Because there are only five latent mean differences, the additional parsimony associated with these models is not substantial in comparison with the corresponding models that do not constrain latent mean differences to be invariant. In each case, the fit of models positing no latent mean differences is at least marginally poorer than the corresponding models in which latent mean differences are freely estimated: Differences in CFI (.005 to .006) and TLI (.006 to .007) are based on comparisons of submodels LIM10E and LIM5E, LIM11E and LIM7E, LIM12E and LIM8E, and LIM13E and LIM9E. However, support for systematic differences in latent means is only marginal.

Because evaluation of latent means is a central, a priori feature of these models, we present mean differences for each of the 28 models that result in mean differences (see Table 7) rather than rely exclusively on indices of fit—especially given that the results based on the fit indices do not seem conclusive. There is a remarkably similar pattern to the mean differences. The set of 378 PSIs between all possible pairs of profiles vary from .993 to over .999 (mean PSI = .998). There are, however, small but systematic differences in the size of means based on complete and partial invariance constraints. In each case the absolute value of mean differences based on complete invariance models is slightly larger than that based on partial invariance. Thus, for example, the standardized mean values for Neuroticism decline about .23 over time for models of complete invariance but only about .20 for models with partial invariance. For Agreeableness, there is an increase of about .30 for models of complete invariance but increases of only about .26 for models of partial invariance. There are smaller increases in Openness and Conscientiousness that are also slightly larger for models with complete invariance. Only for measures of Extraversion are the standardized mean differences consistently close to zero (statistically nonsignificant).

The changes in these latent mean differences over time—especially the decrease in Neuroticism and the increases in Agreeableness, Openness, and Conscientiousness—are consistent with the maturity principle (Caspi et al., 2005) discussed earlier. Indeed,

Table 5
Summary of Goodness-of-Fit Statistics for All Longitudinal Invariance (IN) Models (Time 1/Time 2 Data)

Model and description	χ^2	*df*	CFI	TLI	NFParm	RMSEA
LIM1 (configural IN)						
LIM1: no IN (configural IN)	22,586	6535	.737	.712	845	.040
LIM1A: LIM1 with 60 CWCUs	13,439	6475	.886	.874	905	.026
LIM1B: LIM1 with 57 WWCUs (free)[a]	19,608	6421	.784	.760	959	.036
LIM1C: LIM1 with 57 WWCUs (IN)[a]	19,689	6478	.783	.761	902	.036
LIM1D: LIM1 with 60 CWCUs & 57 WWCUs (free)	11,700	6361	.912	.902	1019	.023
LIM1E: LIM1 with 60 CWCUs & 57 WWCUs (IN)	11,775	6418	.912	.902	962	.023
LIM2 (FL; weak factorial/measurement IN)						
LIM2: IN = FL (weak factorial/measurement IN)	23,310	6810	.729	.716	570	.039
LIM2A: LIM2 with 60 CWCUs	14,031	6750	.881	.874	630	.026
LIM2B: LIM2 with 57 WWCUs (free)[a]	20,277	6696	.777	.763	684	.036
LIM2C: LIM2 with 57 WWCUs (IN)[a]	20,373	6753	.777	.764	627	.036
LIM2D: LIM2 with 60 CWCUs & 57 WWCUs (free)	12,269	6636	.908	.901	744	.023
LIM2E: LIM2 with 60 CWCUs & 57 WWCUs (IN)	12,363	6693	.907	.901	687	.023
LIM3 (FL & Uniq)						
LIM3: IN = FL, Uniq	23,618	6870	.725	.715	510	.039
LIM3A: LIM3 with 60 CWCUs	14,341	6810	.877	.871	570	.027
LIM3B: LIM3 with 57 WWCUs (free)[a]	20,544	6756	.774	.761	624	.036
LIM3C: LIM3 with 57 WWCUs (IN)[a]	20,707	6713	.772	.761	567	.036
LIM3D: LIM3 with 60 CWCUs & 57 WWCUs (free)	12,543	6696	.904	.898	684	.024
LIM3E: LIM3 with 60 CWCUs & 57 WWCUs (IN)	12,695	6753	.903	.897	627	.024
LIM4 (FL & FVCV)						
LIM4: IN = FL, FVCV	23,351	6825	.729	.717	555	.039
LIM4A: LIM4 with 60 CWCUs	14,069	6765	.880	.874	615	.026
LIM4B: LIM4 with 57 WWCUs (free)[a]	20,309	6711	.777	.763	684	.036
LIM4C: LIM4 with 57 WWCUs (IN)[a]	20,407	6768	.776	.764	612	.036
LIM4D: LIM4 with 60 CWCUs & 57 WWCUs (free)	12,298	6651	.907	.901	729	.023
LIM4E: LIM4 with 60 CWCUs & 57 WWCUs (IN)	12,393	6708	.907	.901	672	.023
LIM5 (FL & Inter; strong factorial/measurement IN)						
LIM5D: IN = FL, Inter, with 60 CWCUs & 57 WWCUs (free)	12,796	6691	.900	.893	689	.024
LIM5E: LIM5D with 60 CWCUs & 57 WWCUs (IN)	12,888	6748	.899	.893	632	.024
LIM5Dp: LIM5D with Inter (P-IN), 60 CWCUs & 57 WWCUs (free)	12,524	6680	.904	.898	700	.024
LIM5Ep: LIM5D with Inter (P-IN), with 60 CWCUs & 57 WWCUs (IN)	12,619	6737	.904	.898	643	.024
LIM6 (FL, FVCV, Uniq)						
LIM6D: IN = FL, FVCV, Uniq, with 60 CWCUs & 57 WWCUs (free)	12,578	6711	.904	.898	669	.024
LIM6E: LIM6D with 60 CWCUs & 57 WWCUs (IN)	12,729	6768	.901	.897	612	.024
LIM7 (FL, Uniq, Inter; strict factorial/measurement IN)						
LIM7D: IN = FL, Uniq, Inter, with 60 CWCUs & 57 WWCUs (free)	13,070	6751	.896	.890	629	.024
LIM7E: LIM7D with 60 CWCUs & 57 WWCUs (IN)	13,222	6808	.895	.890	572	.024
LIM7Dp: LIM7D with Inter (P-IN), with 60 CWCUs & 57 WWCUs (free)	12,799	6740	.901	.895	640	.024
LIM7Ep: LIM7D with Inter (P-IN), with 60 CWCUs & 57 WWCUs (IN)	12,950	6797	.899	.894	583	.024
LIM8 (FL, FVCV, Inter)						
LIM8D: IN = FL, FVCV, Inter, with 60 CWCUs & 57 WWCUs (free)	12,826	6706	.900	.893	674	.024
LIM8E: LIM8D with 60 CWCUs & 57 WWCUs (IN)	12,919	6763	.899	.893	617	.024
LIM8Dp: LIM8D with Inter (P-IN), with 60 CWCUs & 57 WWCUs (free)	12,554	6695	.904	.898	685	.024
LIM8Ep: LIM8D with Inter (P-IN), with 60 CWCUs & 57 WWCUs (IN)	12,649	6752	.903	.898	628	.024
LIM9 (FL, Uniq, FVCV, Inter)						
LIM9D: IN = FL, FVCV, Uniq, Inter, with 60 CWCUs & 57 WWCUs (free)	13,106	6766	.896	.890	614	.024
LIM9E: LIM9D with 60 CWCUs & 57 WWCUs (IN)	13,257	6823	.894	.890	557	.025
LIM9Dp: LIM9D with Inter (P-IN), with 60 CWCUs & 57 WWCUs (free)	12,834	6755	.900	.895	625	.024
LIM9Ep: LIM9D with Inter (P-IN), with 60 CWCUs & 57 WWCUs (IN)	12,985	6812	.899	.894	568	.024
LIM10 (FL, Inter, LFMn; latent mean IN)						
LIM10D: IN = FL, Inter, LFMn, with 60 CWCUs & 57 WWCUs (free)	13,166	6696	.894	.887	684	.025
LIM10E: LIM10D with 60 CWCUs & 57 WWCUs (IN)	13,258	6753	.893	.887	627	.025

Table 5 (*continued*)

Model and description	χ^2	*df*	CFI	TLI	NFParm	RMSEA
LIM10Dp: LIM10D with Inter (P-IN), with 60 CWCUs & 57 WWCUs (free)	12,765	6685	.900	.894	695	.024
LIM10Ep: LIM10D with Inter (P-IN), 60 CWCUs & 57 WWCUs (IN)	12,859	6742	.900	.894	638	.024
LIM11 (FL, Uniq, Inter, LFMn; manifest mean IN)						
LIM11D: IN = FL, Uniq, Inter, LFMn, with 60 CWCUs & 57 WWCUs (free)	13,440	6756	.890	.884	624	.025
LIM11E: LIM11D with 60 CWCUs & 57 WWCUs (IN)	13,593	6813	.889	.883	567	.025
LIM11Dp: LIM11D with Inter (P-IN), 60 CWCUs & 57 WWCUs (free)	13,039	6745	.897	.891	635	.024
LIM11Ep: LIM11D with Inter (P-IN), 60 CWCUs & 57 WWCUs (IN)	13,191	6802	.895	.890	578	.024
LIM12 (FL, FVCV, Inter, LFMn)						
LIM12D: IN = FL, FVCV, Inter, LFMn, with 60 CWCUs & 57 WWCUs (free)	13,196	6711	.894	.887	669	.025
LIM12E: LIM12D with 60 CWCUs & 57 WWCUs (IN)	13,289	6768	.893	.887	612	.025
LIM12Dp: LIM12D with Inter (P-IN), 60 CWCUs & 57 WWCUs (free)	12,794	6700	.900	.894	680	.024
LIM12Ep: LIM12D with Inter (P-IN), 60 CWCUs & 57 WWCUs (IN)	12,889	6757	.899	.894	623	.025
LIM13 (FL, Uniq, FVCV, Inter, LFMn; complete factorial IN)						
LIM13D: IN = FL, Uniq, FVCV, Inter, LFMn, with 60 CWCUs & 57 WWCUs (free)	13,476	6771	.890	.884	609	.025
LIM13E: LIM13D with 60 CWCUs & 57 WWCUs (IN)	13,628	6828	.889	.883	552	.025
LIM13Dp: LIM13D with Inter (P-IN), 60 CWCUs & 57 WWCUs (free)	13,074	6817	.896	.891	620	.024
LIM13Ep: LIM13D with Inter (P-IN), 60 CWCUs & 57 WWCUs (IN)	13,226	6817	.895	.890	563	.024

Note. For multiple-group IN models, IN refers to the sets of parameters constrained to be invariant across the multiple groups. The *p* in model names (e.g., LIM5Dp) indicates partial IN (P-IN). CFI = comparative fit index; TLI = Tucker–Lewis index; NFParm = number of free parameters; RMSEA = root-mean-square error of approximation; LIM = longitudinal IN model; CWCUs = cross-wave correlated uniquenesses (CUs); WWCUs = within-wave CUs; FL = factor loadings; Uniq = item uniquenesses; FVCV = factor variances–covariances; Inter = item intercepts; LFMn = latent factor means.
[a] Model results in improper solutions and should be interpreted cautiously (or ignored).

given the relatively short interval between the two measures, it might be surprising that the differences are as large as they are. However, it is also important to note that these results are based on responses by the same students in their final year of high school and again several years later, a period during which changes in maturity might be expected to be significant.

Discussion

Summary and Implications

The a priori Big Five factors are clearly identified by both ESEM and ICM-CFA. The pattern and even the sizes of factor loadings are similar for the two approaches. However, the ESEM solution fits the data much better than does the ICM-CFA solution and resulted in substantially less correlated factors (*Mdn* absolute r = .06 vs. .20) that are consistent with Big Five theory.

Subsequent ESEM analyses support measurement invariance over gender and over time—analyses that could not have been done appropriately with traditional EFA approaches (or ICM-CFA models that were not able to fit the data). The gender invariance analysis showed that women scored higher on all five NEO-FFI factors, whereas the analysis of test–retest data was supportive of the maturity principle (Caspi et al., 2005). Although consistent with previous research based on manifest variables, this is apparently the first research to even pursue these issues in relation to latent Big Five factors and appropriate tests of full measurement and structural invariance in relation to a detailed taxonomy of invariance models (e.g., see Table 1). This is critical in that measurement invariance assumptions are prerequisite to making valid mean comparisons—particularly the assumption of strong measurement invariance with full or at least partial invariance of item intercepts. Whereas we focused on mean differences across gender and over time, strong measurement invariance requirements are equally relevant to all Big Five studies of mean differences for other groups or relations with other constructs. More generally, we recommend that subsequent CFA studies routinely consider ESEM solutions as a viable alternative, even when the fit of CFA solutions is apparently acceptable.

Strengths, Limitations, and Directions for Further Research

The size of factor correlations. Big Five factors are posited to be relatively uncorrelated. This was a key issue in the McCrae et al. (1996; also see Parker et al., 1993) criticism of CFA, because they suggested that forcing an ICM-CFA structure would lead to inflated correlations among the Big Five factors. Our results support this contention. 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 and the differences can be substantial (e.g., .34 vs. .72; Marsh et al., 2009). This issue is also relevant to research based on simple scale scores and EFA factor scores. Correlations based on (a) ICM-CFA latent factors are likely

Table 6
Test–Retest Correlations Between Matching Big Five Factors at T1 and T2

Model	T2 NEUR with T1 NEUR	T2 EXTR with T1 EXTR	T2 OPEN with T1 OPEN	T2 AGRE with T1 AGRE	T2 CONC with T1 CONC
1					
LIM1	.777	.888	.915	.812	.899
LIM1A	.749	.812	.849	.767	.810
LIM1B[a]	.793	.917	.953	.831	.931
LIM1C[a]	.794	.917	.953	.830	.932
LIM1D	.760	.841	.873	.784	.833
LIM1E	.760	.842	.874	.783	.835
2					
LIM2	.780	.892	.921	.811	.901
LIM2A	.750	.813	.850	.767	.811
LIM2B[a]	.794	.920	.957	.830	.933
LIM2C[a]	.795	.920	.957	.830	.934
LIM2D	.759	.843	.880	.781	.831
LIM2E	.760	.843	.879	.780	.832
3					
LIM3	.780	.892	.922	.810	.900
LIM3A	.748	.818	.850	.764	.814
LIM3B[a]	.794	.920	.959	.829	.932
LIM3C[a]	.795	.920	.958	.828	.933
LIM3D	.756	.842	.878	.880	.831
LIM3E	.757	.843	.879	.881	.832
4					
LIM4	.775	.891	.921	.810	.901
LIM4A	.744	.817	.850	.764	.814
LIM4B[a]	.791	.920	.957	.829	.933
LIM4C[a]	.791	.920	.958	.829	.935
LIM4D	.756	.842	.878	.780	.831
LIM4E	.756	.842	.878	.783	.831

Note. For a description of the models tested (e.g., LIM1, LIM1A) and their fit to the data, see Table 5. T1 = Time 1; T2 = Time 2; NEUR = Neuroticism; EXTR = Extraversion; OPEN = Openness; AGRE = Agreeableness; CONC = Conscientiousness; LIM = longitudinal invariance model.
[a] These models result in improper solutions and should be interpreted cautiously (or ignored).

to be inflated as shown here; (b) EFA factor scores are likely to be attenuated (because they do not correct for unreliability); and (c) manifest scale scores are likely to be both inflated and attenuated (although it would be difficult to determine the relative sizes of these counterbalancing biases). In all ICM-CFA applications, factor correlations will be at least somewhat inflated unless all nontarget loadings are close to zero. This results in multicollinearity and undermines discriminant validity in relation to predicting other outcomes and providing distinct profiles of personality.

Complex measurement error structures. Big Five research has largely ignored fundamental issues related to complex structures of measurement error. Although Big Five researchers routinely report coefficient alpha estimates of reliability, the "state of the art" has moved well beyond these historically acceptable measures. Coefficient alpha estimates of reliability provide an index of one aspect of measurement error, but they largely ignore other aspects of unreliability and do not correct parameter estimates for unreliability (also see Sijtsma, 2009). Particularly in path models with many different constructs, the failure to control for measurement error can have unanticipated results (see discussion of the *phantom* effect by Marsh et al., 2010).

For test–retest correlations, there are at least two crucial components of measurement error that are typically off-setting at least to an extent. Measurement error for constructs at each time considered separately attenuates the sizes of correlations. However, responses to the same items on two occasions are typically more positively correlated than can be explained in terms of correlations between the factors that they represent, and this inflates the correlations. Indeed, typical short-term test–retest correlations in Big Five studies (.86–.90; McCrae & Costa, 2004) corrected for typical internal consistency reliability estimates (.68–.86; Costa & McCrae, 1992) would result in test–retest correlations greater than 1.0. In the present investigation, estimated test–retest correlations were based on a longer time interval but still approached 1.0 and resulted in improper solutions. Problems such as these led Marsh and Hau (1996) to recommend that CUs always be incorporated into evaluations of test–retest correlations. Here we demonstrated how this can be done in ESEM models.

In the present investigation, we also evaluated an additional source of measurement error that is idiosyncratic to the design of the NEO-FFI. More specifically, we posited that items from the same facet of the long form of the NEO would be more highly correlated than would items designed to measure the same factor but from different facets. There was strong support for this additional source of measurement error; inclusion of WWCUs contributed substantially to goodness of fit, and they were reasonably invariant over responses by men and women and over time for responses by the same individuals. Although these WWCUs were idiosyncratic to the design of the NEO-FFI, there are many other method effects that also distort findings if not controlled. Indeed,

Table 7
Patterns of Mean Differences Over Time for Big Five Factors

Model and description	NEUR	EXTR	OPEN	AGRE	CONC
LIM5 (FL & Inter; strong factorial/measurement IN)					
LIM5D: IN = FL, Inter, with 60 CWCUs & 57 WWCUs (free)	−.228	.015	.176	.332	.227
LIM5E: LIM5D with 60 CWCUs & 57 WWCUs (IN)	−.226	.016	.175	.331	.226
LIM5Dp: LIM5D with Inter (P-IN), with 60 CWCUs & 57 WWCUs (free)	−.202	.032	.127	.260	.194
LIM5Ep: LIM5D with Inter (P-IN), with 60 CWCUs & 57 WWCUs (IN)	−.200	.033	.127	.259	.193
LIM7 (FL, Uniq, Inter; strict factorial/measurement IN)					
LIM7D: IN = FL, Uniq, Inter, with 60 CWCUs & 57 WWCUs (free)	−.227	.015	.178	.336	.226
LIM7E: LIM7D with 60 CWCUs & 57 WWCUs (IN)	−.227	.015	.178	.334	.226
LIM7Dp: LIM7D with Inter (P-IN), with 60 CWCUs & 57 WWCUs (free)	−.203	.032	.129	.262	.193
LIM7Ep: LIM7D with Inter (P-IN), with 60 CWCUs & 57 WWCUs (IN)	−.201	.032	.128	.261	.193
LIM8 (FL, FVCV, Inter)					
LIM8D: IN = FL, FVCV, Inter, with 60 CWCUs & 57 WWCUs (free)	−.235	.014	.171	.336	.227
LIM8E: LIM8D with 60 CWCUs & 57 WWCUs (IN)	−.235	.014	.171	.336	.227
LIM8Dp: LIM8D with Inter (P-IN), with 60 CWCUs & 57 WWCUs (free)	−.209	.032	.123	.255	.195
LIM8Ep: LIM8D with Inter (P-IN), with 60 CWCUs & 57 WWCUs (IN)	−.208	.032	.123	.255	.195
LIM9 (FL, Uniq, FVCV, Inter)					
LIM9D: IN = FL, FVCV, Uniq, Inter, with 60 CWCUs & 57 WWCUs (free)	−.234	.014	.171	.337	.226
LIM9E: LIM9D with 60 CWCUs & 57 WWCUs (IN)	−.235	.014	.171	.326	.227
LIM9Dp: LIM9D with Inter (P-IN), with 60 CWCUs & 57 WWCUs (free)	−.220	.031	.123	.255	.194
LIM9Ep: LIM9D with Inter (P-IN), with 60 CWCUs & 57 WWCUs (IN)	−.209	.031	.123	.255	.194

Note. See Tables 1 and 5 for a description of the models. Each of the 16 models provides estimates of latent mean differences over time for the Big Five factors under different assumptions. The pattern of differences across the 16 models is very similar, with correlations varying from .993 to .999 (mean r = .9975). The *p* in model names (e.g., LIM5Dp) indicates partial IN (P-IN). NEUR = Neuroticism; EXTR = Extraversion; OPEN = Openness; AGRE = Agreeableness; CONC = Conscientiousness; LIM = longitudinal IN model; FL = factor loadings; Inter = item intercepts; IN = invariance (for multiple-group IN models, IN refers to the sets of parameters constrained to be invariant across the multiple groups); CWCUs = cross-wave correlated uniquenesses (CUs); WWCUs = within-wave CUs; Uniq = item uniquenesses; FVCV = factor variances–covariances.

the logic underlying these WWCUs is similar to that based on the CWCUs that are routinely incorporated into longitudinal models.

Taxonomy of measurement invariance models. In psychological assessment research, there has been an unfortunate schism between factor analysts, who focus on the invariance of factor structures over groups or over time, and measurement invariance researchers, who focus on differential item functioning and assumptions underlying the appropriate comparison of latent or manifest mean test scores. The taxonomy of invariance models proposed here (also see Marsh et al., 2009) brings together these two approaches. The actual models would be equally appropriate for either ESEM or CFA approaches, although not for traditional approaches to EFA. Although the taxonomy incorporates a richer selection of models, it is not meant to be exhaustive. Indeed, here we expanded the basic taxonomy to include diverse variations of the models to incorporate a priori CUs and ex post facto partial invariance. Furthermore, researchers might choose to focus on some models rather than others in a specific application.

Our taxonomy is more comprehensive than traditional approaches to measurement invariance, allowing us to integrate concerns typically not considered in studies of measurement invariance. Tests of measurement invariance typically follow a particular sequence of tests in which the fulfillment of invariance at each step is dependent upon fulfillment of invariance on the previous step. However, there is no reason why an applied researcher, for example, should not evaluate the invariance of item uniquenesses even if there is not support for the invariance of item intercepts. Indeed, this is routine practice in tests from the factor analysis perspective that typically do not even consider item-intercept invariance. Furthermore, tests of measurement invariance typically do not consider the invariance of variance–covariance matrices, so that it is unclear where they would fit into a measurement invariance sequence. In addition, tests of measurement invariance base critical decisions (e.g., invariance of item intercepts) on the comparison of one pair of models. In contrast, our approach provides tests of the invariance of the same set of parameter estimates based on many different pairs of models. Although this feature of our taxonomy appears to be potentially valuable, more research is needed to evaluate this difference. Finally, it is important to emphasize that we use terms such as *configural invariance, weak invariance, strong invariance,* and *strict invariance* in precisely the same way as these terms are traditionally used in tests of measurement invariance and that we use the same models as used in tests of measurement invariance.

In summary, we suggest that this taxonomy makes two main contributions. First, it provides a concrete set of models that incorporates all or most of the specific models considered by both factor analysts and measurement invariance researchers and then identifies an apparent limitation in much personality research. Second, the application of this taxonomy demonstrates the flexibility of the ESEM approach, which integrates many of the best features of traditional CFA and EFA approaches.

Goodness of fit. Quantitative psychologists are constantly seeking universal "golden rules"—guidelines that allow them to make objective interpretations of their data rather than being forced to defend subjective interpretations (Marsh et al., 2004). Marsh et al. (2004) likened this to pursuit of the mythical Golden Fleece, the Fountain of Youth, and absolute truth and beauty—appealing, but unlikely to ever be realized. Over time a plethora of different indices have been proposed; most were substantially related but had somewhat different properties (e.g., Marsh et al., 1988). However, there is even less consensus today than in the past as to what constitutes an acceptable fit; some still treat the indices and recommended cutoffs as golden rules, others argue that fit indices should be discarded altogether, a few argue that we should dispense with multiple indicators altogether and rely solely on chi-square goodness-of-fit indices, and many (like us) argue that they should be treated as rough guidelines to be interpreted cautiously in combination with other features of the data (see the special issue of *Personality and Individual Differences* [Vernon & Eysenck, 2007], in which different authors advocate each of these positions). These problems are not resolved by comparing the fit of alternative models, because applied researchers are still left with the task of deciding whether differences in model fit are sufficiently large to be substantively meaningful. Nevertheless, there are important advantages in using an a priori taxonomy of models that facilitates communication and allows the researchers to pinpoint sources of misfit.

Given the lack of consensus about the appropriate use of fit indices, it is not surprising that there is also ambiguity in their application in ESEM and to the new issues that ESEM raises. For example, because the number of factor loadings alone in ESEM applications is the product of the number of items times the number of factors, the total number of parameter estimates in ESEM applications can be massively more than in the typical CFA application. This feature might make problematic any index that does not control for model parsimony (due to capitalization on chance) and calls into question the appropriateness of controls for parsimony in those indices that do. In the present investigation (with 60 items and five factors) interpretations based on CFI and TLI in relation to existing standards were reasonably interpretable, whereas almost all the models considered here were "excellent" in relation to an RMSEA cutoff of .05. Although changes in RMSEA values for nested models behaved more reasonably, even here there was not good differentiation between models for which the fit was apparently relatively good and those for which it was relatively poor.

In summary, the introduction of ESEM provides no panacea to evaluating goodness of fit. Clearly there is need for more research—particularly in relation to applied practice for which ESEM is likely to be most beneficial. However, given the current thinking about goodness of fit in CFA applications, unambiguous cutoff values of acceptable fit—or even differences in fit for nested models—seem unlikely. In the meantime, we suggest that applied researchers use an eclectic approach based on a subjective integration of a variety of different indices, detailed evaluations of the actual parameter estimates in relation to theory, a priori predictions, common sense, and a comparison of viable alternative models specifically designed to evaluate goodness of fit in relation to key issues. This is consistent with the approach we used here (and incorporates an emphasis on the careful consideration of parameter estimates that constitutes best practice in personality research based on EFAs). In particular, we recommend that cutoff values for goodness-of-fit indices be interpreted cautiously and not used mindlessly.

Other alternative approaches based on item aggregates. Other researchers have used a variety of different strategies that allowed them to apply CFA approaches to NEO responses (or other Big Five measures). However, most of these Big Five studies were not based on analyses at the item level, instead using one of a variety of aggregate scores based on the mean response to different items; for example, facet scores (e.g., McCrae et al., 1996; Saucier, 1998; Small et al., 2003), parcel scores (e.g., Allemand et al., 2008, 2007; Lüdtke et al., 2009; Marsh, Trautwein, et al., 2006), or scale scores (e.g., Mroczek & Spiro, 2003). Although potentially appropriate and useful for some specific purposes, aggregate scores have important limitations to their use. Thus, for example, the use of item aggregates instead of individual items would not allow researchers to evaluate (at the level of the individual item) differential item functioning, items with low target factor loadings, or items with substantial nontarget cross-loadings (or modification indices that are indicative of cross-loadings). Analyses of item aggregates also mask potential method effects that are idiosyncratic to specific items.

Particularly when there are substantial cross-loadings at the individual item level, analyses based on item aggregates that mask these effects are likely to result in inflated factor correlations in much the same way as the ICM-CFA models resulted in inflated factor correlations compared with those calculated with the ESEM approach. Also, results based on analyses of item aggregates would not provide unambiguous information on how existing instruments should be improved by identifying potentially weak items. Furthermore, it is well known (see Marsh, 2007) that the use of item parcels typically results in systematically inflated factor loadings, lower indicator uniquenesses, and inflated goodness-of-fit indices relative to corresponding analyses at the individual item level. Thus, results about the quality of the factor solution based on item aggregates are not comparable to those based on individual items. Of particular relevance to the present investigation, it is unlikely that tests of measurement invariance based on our taxonomy of invariance models would be valid unless they are based on responses to individual items.

Although a detailed discussion of the rationale for using item aggregates is beyond the scope of the present investigation (see Little, Cunningham, Shahar, & Widaman, 2002; Marsh, 2007), most of these rationales are based at least implicitly on the assumption that the a priori model tested at the item level provides a good fit to the data. However, it is difficult to support this argument unless analyses are actually done at the item level. Nevertheless, the controversial literature on the appropriate use of item aggregates does suggest some special cases in which the use of item aggregates might be justified (e.g., when the sample size is small or the number of items is large). Our position is not that analyses based on item aggregates are inherently bad but merely that results should be interpreted appropriately and with due caution. We suspect that some analyses based on item aggregates were conducted because of problems associated with application of the ICM-CFA approach at the item level so that the ESEM approach demonstrated here provides a viable alternative. Hence, we recommend that applied researchers who choose to do CFA analyses at the item-aggregate level evaluate the appropriateness of the ESEM approach for analyses at the individual item level and compare results on the basis of the two approaches.

Partial invariance based on ex post facto modification indices. In the earlier discussion we indicated that this was an area of concern, a limitation in the present investigation, and a direction for further research. In the present investigation, support for the full invariance of item intercepts in relation to time was marginal and was clearly lacking in relation to gender. We had a choice, as is likely to be the case in many applied studies. We could have adopted a purist perspective and simply not pursued any further analyses. Instead we took a pragmatic perspective and sought support for partial invariance. Although clearly ex post facto, there are several justifications for our decision. First, the sample size was sufficiently large that capitalization on chance was not a major concern. Second, there were 12 items for each Big Five factor, so at least five or six items per factor had invariant intercepts in our tests of gender invariance (and even more for tests of invariance over time). This is very different from many applications, which are based on only a few items per factor such that there may be only one or two items with invariant intercepts after introducing partial invariance. Third, these ex post facto modifications were reasonably invariant over gender and over time, supporting their generalizability and stability. Finally, the patterns of gender differences and latent mean differences over time were similar for the fully and partially invariant solutions. A stronger approach might be to posit a priori those item intercepts for which invariance is likely to fail or, perhaps, to evaluate the ex post facto reasonableness of item intercepts that were not invariant (e.g., Roberts et al., 2006). However, we had no a priori basis for knowing which item intercepts would fail and suspect that this is likely to be the case for most applied studies. Also, we have always been a bit suspect of the reasonableness of ex post facto explanations (if they are so reasonable, then why was the explanation not an a priori hypothesis). Furthermore, such ex post facto scrutiny is likely to be more valuable when constructing an instrument—the stage when applied researchers are selecting the best items from a large pool of items—than when evaluating one of the most widely used instruments in personality research. Nevertheless, we readily concede that this issue is a limitation in our study and one that needs further research and consideration in the context of ESEM and measurement invariance more generally.

Conclusions

Why have Big Five researchers not taken more advantage of the tremendous advances in statistical methodology that appear to be highly relevant to important substantive concerns such as those considered here? Many of these advances are based substantially on CFA and related statistical techniques. We argued here that the traditional ICM-CFA model is not appropriate for the NEO-FFI and suspect that this would also be the case for many personality measures. Indeed, this is commonly expressed by Big Five researchers (e.g., McCrae et al., 1996) and is consistent with the failure of Big Five CFAs based on item-level responses for NEO instruments to achieve acceptable levels of fit (but see Benet-Martinez & John, 1998). However, personality researchers proclaiming the inappropriateness of CFA are also forced to forgo the many methodological advances that are associated with CFA, an ironic situation in a discipline that has made such extensive use of factor analysis. We suspect that the failure to incorporate these important advances can be overcome—at least to some extent—through application of the ESEM approach, as demonstrated here.

Importantly, the analytical strategies demonstrated here could also be applied in traditional ICM-CFA studies. In this respect, we present the ESEM model as a viable alternative to the ICM-CFA model but do not argue that the ESEM approach should replace the corresponding CFA approach. Indeed, when the more parsimonious ICM-CFA model fits the data as well as the ESEM model does, then the ICM-CFA should be used. However, when the ICM-CFA model is unable to fit the data whereas the ESEM model is able to do so, we suggest that advanced statistical strategies such as those demonstrated here are more appropriately conducted with ESEM models than with ICM-CFA models. From this perspective, our results provide clear evidence that an ESEM approach is more appropriate than a traditional ICM-CFA approach for Big Five responses to the NEO-FFI. Although ESEM is not a panacea and may not be appropriate in some applications, it provides applied personality researchers with considerable flexibility in addressing issues such as those raised here when the traditional ICM-CFA approach is not appropriate. Because ESEM is a new statistical tool, best practice will have to evolve with experience. Nevertheless, results of the present investigation (also see Marsh et al., 2009) provide considerable promise for the application of ESEM in Big Five studies and in psychological assessment research more generally.

References

Allemand, M., Zimprich, D., & Hendriks, A. A. J. (2008). Age differences in five personality domains across the life span. *Developmental Psychology, 44,* 758–770.

Allemand, M., Zimprich, D., & Hertzog, C. (2007). Cross-sectional age differences and longitudinal age changes of personality in middle adulthood and old age. *Journal of Personality, 75,* 323–358.

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

Benet-Martinez, 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.

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

Borkenau, P., & Ostendorf, F. (1993). *NEO-Fünf-Faktoren Inventar nach Costa und McCrae.* [NEO Five Factor Inventory after Costa and McCrae]. Göttingen, Germany: Hogrefe.

Boyle, G. J. (2008). Critique of the five-factor model of personality. In G. J. Boyle, G. Matthews, & D. H. Saklofske (Eds.), *The SAGE Handbook of Personality Theory and Assessment: Vol. 1. Personality theories and models* (pp. 295–312). Thousand Oaks, CA: Sage.

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.

Caprara, G. V., & Perugini, M. (1994). Personality described by adjectives: The generalizability of the Big Five to the Italian lexical context. *European Journal of Personality, 8,* 357–369.

Caspi, A., Roberts, B. W., & Shiner, R. L. (2005). Personality development: Stability and change. *Annual Review of Psychology, 56,* 453–484.

Caspi, A., & Shiner, R. L. (2006). Personality development. In N. Eisenberg, W. Damon, & R. M. Lerner (Eds.), *Handbook of Child Psychology: Vol. 3. Social, emotional, and personality development* (6th ed., pp. 300–365). Hoboken, NJ: Wiley.

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

Cheung, G. W., & Rensvold, R. B. (2001). The effects of model parsimony and sampling error on the fit of structural equation models. *Organizational Research Methods, 4,* 236–264.

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.

Cliff, N. (1966). Orthogonal rotation to congruence. *Psychometrika, 31,* 33–42.

Costa, P. T., Jr., & McCrae, R. R. (1989). *The NEO-PI/NEO-FFI manual supplement.* Odessa, FL: Psychological Assessment Resources.

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

Costa, P. T., Jr., & McCrae, R. R. (1995). Domains and facets: Hierarchical personality assessment using the Revised NEO Personality Inventory. *Journal of Personality Assessment, 64,* 21–50.

Costa, P. T., Jr., Terracciano, A., & McCrae, R. R. (2001). Gender differences in personality traits across cultures: Robust and surprising findings. *Journal of Personality and Social Psychology, 81,* 322–331.

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.

Donnellan, M. B., & Lucas, R. E. (2008). Age differences in the Big Five across the life span: Evidence from two national samples. *Psychology and Aging, 23,* 558–566.

Feingold, A. (1994). Gender differences in personality: A meta-analysis. *Psychological Bulletin, 116,* 429–456.

Fraley, R., & Roberts, B. W. (2005). Patterns of continuity: A dynamic model for conceptualizing the stability of individual differences in psychological constructs across the life course. *Psychological Review, 112,* 60–74.

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

Goldberg, L. R. (1990). An alternative "description of personality": The Big Five factor structure. *Journal of Personality and Social Psychology, 59,* 1216–1229.

Gosling, S. D., Rentfrow, P. J., & Swann, W. B., Jr. (2003). A very brief measure of the Big Five personality domains. *Journal of Research in Personality, 37,* 504–528.

Gustavsson, J. P., Eriksson, A. K., Hilding, A., Gunnarsson, M., & Ostensson, C. G. (2008). Measurement invariance of personality traits from a five-factor model perspective: Multi-group confirmatory factor analyses of the HP5 inventory. *Scandinavian Journal of Psychology, 49,* 459–467.

Hall, J. A. (1984). *Nonverbal sex differences: Communication accuracy and expressive style.* Baltimore, MD: Johns Hopkins University Press.

Holden, R. R., & Fekken, G. C. (1994). The NEO Five-Factor Inventory in a Canadian context: Psychometric properties for a sample of university women. *Personality and Individual Differences, 17,* 441–444.

Jackson, J. J., Bogg, T., Walton, K., Wood, D., Harms, P. D., Lodi-Smith, J. L., & Roberts, B. W. (2009). Not all conscientiousness scales change alike: A multimethod, multisample study of age differences in the facets of conscientiousness. *Journal of Personality and Social Psychology, 96,* 446–459.

John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (Vol. 2, pp. 102–138). New York, NY: Guilford Press.

Jöreskog, K. G. (1979). Statistical estimation of structural models in longitudinal investigations. In J. R. Nesselroade & B. Baltes (Eds.), *Longitudinal research in the study of behavior and development* (pp. 303–351). New York, NY: Academic Press.

Jöreskog, K. G., & Sörbom, D. (1988). *LISREL 7: A guide to the program and applications* (2nd ed.). Chicago, IL: SPSS.

Klimstra, T. A., Hale, W. W., III, Raaijmakers, Q. A. W., Branje, S. J. T., & Meeus, W. H. J. (2009). Maturation of personality in adolescence. *Journal of Personality and Social Psychology, 96,* 898–912.

Köller, O., Watermann, R., Trautwein, U., & Lüdtke, O. (Eds.). (2004). *Wege zur Hochschulreife in Baden-Württemberg. TOSCA—Eine Untersuchung an allgemein bildenden und beruflichen Gymnasien* [Educational pathways to college in Baden-Württemberg: TOSCA—a study of traditional and vocational gymnasium schools]. Opladen, Germany: Leske & Budrich.

Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling, 9,* 151–173.

Lüdtke, O., Trautwein, U., & Husemann, N. (2009). Goal and personality trait development in a transitional period: Assessing change and stability in personality development. *Personality and Social Psychology Bulletin, 35,* 428–441.

Lüdtke, O., Trautwein, U., Nagy, G., & Köller, O. (2004). Eine Validierungsstudie zum NEO-FFI in einer Stichprobe junger Erwachsener: Effekte des Itemformats, faktorielle Validität und Zusammenhänge mit Schulleistungsindikatoren [A validation of the NEO-FFI in a sample of young adults: Effects of the response format, factorial validity, and relations with indicators of academic achievement]. *Diagnostica, 50,* 134–144.

Maccoby, E. E., & Jacklin, C. N. (1974). *The psychology of sex differences.* Stanford, CA: Stanford University Press.

Marsh, H. W. (1994). Confirmatory factor analysis models of factorial invariance: A multifaceted approach. *Structural Equation Modeling, 1,* 5–34.

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., Balla, J. R., & Hau, K. T. (1996). An evaluation of incremental fit indices: A clarification of mathematical and empirical processes. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling techniques* (pp. 315–353). Hillsdale, NJ: Erlbaum.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin, 103,* 391–410.

Marsh, H. W., & Hau, K.-T. (1996). Assessing goodness of fit: Is parsimony always desirable? *Journal of Experimental Education, 64,* 364–390.

Marsh, H. W., & Hau, K.-T. (2007). Applications of latent-variable models in educational psychology: The need for methodological-substantive synergies. *Contemporary Educational Psychology, 32,* 151–171.

Marsh, H. W., Hau, K.-T., Balla, J. R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research, 33,* 181–220.

Marsh, H. W., Hau, K.-T., & Grayson, D. (2005). Goodness of fit evaluation in structural equation modeling. In A. Maydeu-Olivares & J. McCardle (Eds.), *Psychometrics: A festschrift to Roderick P. McDonald* (225–340). Hillsdale, NJ: Erlbaum.

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

Marsh, H. W., Martin, A., & Debus, R. (2001). Individual differences in verbal and math self-perceptions: One factor, two factors, or does it depend on the construct? In R. Riding & S. Rayner (Eds.), *Self-perception: International perspectives on individual differences* (pp. 149–170). Westport, CT: Ablex.

Marsh, H. W., Martin, A. J., & Hau, K.-T. (2006). A multiple method perspective on self-concept research in educational psychology: A construct validity approach. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 441–456). Washington, DC: American Psychological Association.

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., Seaton, M., Kuyper, H., Dumas, F., Huguet, P., Regner, I., . . . Gibbons, F. X. (2010). Phantom behavioral assimilation effects: Systematic biases in social comparison choice studies. *Journal of Personality, 78,* 671–710.

Marsh, H. W., Trautwein, U., Lüdtke, O., Köller, O., & Baumert, J. (2006). Integration of multidimensional self-concept and core personality constructs: Construct validation and relations to well-being and achievement. *Journal of Personality, 74,* 403–455.

McCrae, R. R., & Costa, P. T., Jr. (1989). Rotation to maximize the construct validity of factors in the NEO Personality Inventory. *Multivariate Behavioral Research, 24,* 107–124.

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

McCrae, R. R., & Costa, P. T., Jr. (2004). A contemplated revision of the NEO Five-Factor Inventory. *Personality and Individual Differences, 36,* 587–596.

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.

Meredith, W. (1964). Rotation to achieve factorial invariance. *Psychometrika, 29,* 187–206.

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

Meredith, W., & Teresi, J. (2006). An essay on measurement and factorial invariance. *Medical Care, 44*(Suppl. 3), S69–S77.

Mroczek, D. K., & Spiro, A., III. (2003). Modeling intraindividual change in personality traits: Findings from the normative aging study. *Journal of Gerontology: Psychological Sciences, 58*B, 153–165.

Muthén, L. K., & Muthén, B. (2008). *Mplus user's guide.* Los Angeles CA: Muthén & Muthén.

Nye, C., Roberts, B. W., Saucier, G., & Zhou, X. (2008). Testing the measurement equivalence of personality adjective items across cultures. *Journal of Research in Personality, 42,* 1524–1536.

Parker, J. D. A., Bagby, R. M., & Summerfeldt, L. J. (1993). Confirmatory factor analysis of the Revised Neo-Personality Inventory. *Personality and Individual Differences, 15,* 463–466.

Paunonen, S. V. (2003). Big Five factors of personality and replicated predictions of behavior. *Journal of Personality and Social Psychology, 84,* 411–422.

Paunonen, S. V., & Ashton, M. C. (2001). Big Five factors and facets and the prediction of behavior. *Journal of Personality and Social Psychology, 81,* 524–539.

Reise, S. P., Smith, L. R., & Furr, M. (2001). Invariance on the NEO PI-R Neuroticism Scale. *Multivariate Behavioral Research, 36,* 83–110.

Roberts, B. W., & DelVecchio, W. F. (2000). The rank-order consistency of personality traits from childhood to old age: A quantitative review of longitudinal studies. *Psychological Bulletin, 126,* 5–25.

Roberts, B. W., Walton, K. E., & Viechtbauer, W. (2006). Patterns of mean-level change in personality traits across the life course: A meta-analysis of longitudinal studies. *Psychological Bulletin, 132,* 1–25.

Robins, R. W., Fraley, R. C., Roberts, B. W., & Trzesniewski, K. H. (2001). A longitudinal study of personality change in young adulthood. *Journal of Personality, 69,* 617–640.

Saucier, G. (1998). Replicable item-cluster subcomponents in the NEO Five-Factor Inventory. *Journal of Personality Assessment, 70,* 263–276.

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.

Small, B. J., Hertzog, C., Hultsch, D. F., & Dixon, R. A. (2003). Stability and change in adult personality over 6 years: Findings from the Victoria longitudinal study. *Journal of Gerontology: Psychological Sciences, 58*B, 166–176.

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.

Vernon, T., & Eysenck, S. (Eds.). (2007). Special issue on structural equation modeling. *Personality and Individual Differences, 42*(5).

Received August 14, 2009
Revision received January 7, 2010
Accepted January 7, 2010 ■