This is the final accepted (prepublication) version of the following article published in Developmenal Psychology. This material is copyrighted by the American Psychological Association: http://www.apa.org/about/contact/copyright/seek-permission.aspx.

In compliance with regulations of the American Psychological Association, we note that: ***This article may not exactly replicate the final version published in the APA journal. It is not the copy of record.***

Marsh, H. W., Nagengast, B., & Morin, A. J. S. (2012). Measurement invariance of Big-Five factors over the life span: ESEM tests of gender, age, plasticity, maturity, and La Dolce Vita effects. Developmental Psychology. Advance online publication. doi:10.1037/a0026913.

Measurement Invariance of Big-Five Factors Over the Lifespan: ESEM Tests of Gender, Age, Plasticity, Maturity and La Dolce Vita Effects

3 June 2010
Revised 10 December, 2010
Revised 23 April, 2011

Abstract

In a substantive-methodological synergy, the present investigation applies new and evolving approaches to factor analysis to substantively important developmental issues of how five-factor-approach (FFA) personality measures vary with gender, age, and their interaction. Confirmatory factor analyses (CFA) conducted at the item-level often do not support a priori FFA structures, due in part to the overly restrictive assumptions of CFA models. Here we demonstrate that exploratory structural equation modeling (ESEM), an integration of CFA and EFA, overcomes these problems with the 15-item Big Five Inventory administered as part of the nationally representative British Household Panel Study (n= 14,021; age: 15-99, Mage=47.1). ESEM fit the data substantially better and resulted in much more differentiated (less correlated) factors than CFA. Methodologically, we extend ESEM (introducing ESEM-Within-CFA models and a hybrid of multiple groups and multiple indicators multiple causes – MIMIC – models), evaluating full measurement invariance and latent mean differences over age, gender, and their interaction. Substantively the results showed that women had higher latent scores for all big-five factors except for Openness and that these gender differences were consistent over the entire lifespan. Substantial non-linear age effects led to the rejection of the plaster hypothesis and the maturity principle but did support a newly proposed La Dolce Vita effect in old age. In later years, individuals become happier (more agreeable and less neurotic), more self-content and self-centered (less extroverted and open), more laid back and satisfied with what they have (less conscientious, open, outgoing and extroverted), and less preoccupied with productivity.

This study is a substantive-methodological synergy, bringing to bear new approaches to factor analysis to substantively important developmental issues of how five-factor-approach (FFA) personality measures vary with gender, age, and their interaction. In particular, there has been surprisingly little methodologically rigorous research evaluating changes in FFA personality measures across the lifespan—especially old age.

Factor analysis has been at the heart of the currently dominant approach in personality research that individual differences in adults' personality can universally be organized in terms of five broad trait domains—the five-factor-approach (FFA) to personality: Extraversion (e.g., sociability, gregariousness, level of activity, experience of positive affect); Agreeableness (e.g., altruistic behavior, trust, warmth, kindness); Conscientiousness (e.g., self-control, task-orientation, rule-abiding; Neuroticism (e.g., distress anxiety, anger, depression); Openness (e.g., originality, creativity, and the acceptance of new ideas; for more detail on these factors as used here, see detailed description in the online supplemental materials). Following Block (2010), we use the generic term of FFA that is not specifically aligned to any particular group of researchers or instruments, but acknowledge that some personality researchers – including Block – are critical of the assumption that the self-report FFA factors really do provide an adequate representation of global personality. From this perspective, we emphasize that our focus is on self-report FFA factors – their measurement, analysis, relation to gender and age – from a developmental perspective.

Exploratory factor analyses (EFAs) have consistently identified the FFA factors and an impressive body of empirical research supports their stability and predictive validity across different populations, settings and countries (e.g., McCrae & Costa, 1997) and its circumplex structure (De Raad & Hofstee, 1993). However, confirmatory factor analyses (CFAs) and structural equation models (SEMs) have typically failed to provide clear support for the FFA based on standard measures (e.g., Marsh, Lüdtke et al., 2010; Vassend & Skrondal, 1997).

Problematic FFA results based on CFAs have led some researchers to question the appropriateness of CFA for FFA research (see Borkenau & Ostendorf, 1990; Church & Burke, 1994; McCrae, Zonderman, Costa, Bond & Paunonen, 1996; Parker, Bagby & Summerfeldt, 1993; Vassend & Skrondal, 1997; also see discussions by Dolan, Oort, Stoel, & Wicherts, 2009; Marsh, Lüdtke et al., 2010). In particular, the independent clusters model (ICM) used in CFA studies that require each indicator to load on only one factor may be too restrictive for FFA research. CFA models typically do not provide an adequate fit to the data and lead to positively biased FFA factor correlations that might distort relations with other constructs as well as inducing multicollinearity (see Ashton, Lee, Goldberg & De Vries, 2009; Marsh, Lüdtke et al., 2010). Such concerns have plagued FFA research and promoted leading FFA proponents such as McCrae et al (1996, p.563, also see Church & Burke, 1994; Costa & McCrae, 1992; 1995; McCrae & Costa, 1997; but also see Borsboom, 2006) 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*

Hence, research into the FFA factor structure based on responses to individual items largely continues to rely on EFA (for exceptions, see Benet-Martinez & John, 1998; Dolan et al., 2009; Gustavsson, Eriksson, Hilding, Gunnarsson & Ostensson, 2008; Marsh, Lüdtke et al., 2010; Reise, Smith & Furr, 2001), despite the limitations of traditional applications of EFA in comparison to the multiple advances made in CFA/SEM models over the past decades (e.g., tests of factorial and measurement invariance, differential item functioning, control for complex measurement error structures). Particularly important for the present investigation is the Dolan et al. (2009) study that extended the traditional EFA approach based on responses to NEO-PI-R big-five instrument and foreshadowed the subsequent development of ESEM through the development of an innovative approach to EFA-based multigroup rotation procedure and tests of measurement invariance (also see Hessen, Dolan, & Wicherts, 2006; Marsh, Lüdtke et al., 2010).

Underpinning FFA research into mean differences between groups (e.g., men and women) and relations with other constructs (e.g., age) are methodological assumptions of factorial and measurement invariance that cannot be appropriately evaluated with traditional EFA approaches. Hence, these assumptions have been largely ignored in most substantive research that continues to rely on FFA scale scores (manifest variables) rather than latent constructs in CFA/SEM models. Furthermore, this gap between applied, substantive research and state-of-the-art methodology appears to be increasing (Borsboom, 2006; Marsh & Hau, 2007). Here we outline a new approach allowing for the incorporation of EFA into the SEM-framework – exploratory structural equation modeling (ESEM; Asparouhov & Muthén, 2009; Marsh, Lüdtke et al., 2010; Marsh et al., 2009) – and its extension in ways to enhance further its applicability in developmental research. ESEM and the extensions presented here have the potential to resolve the many dilemmas of factor analysis in FFA research and have 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. We begin with a brief overview of substantive research into gender and age differences in FFA factors and then introduce methodological issues that place limits on this research.

Substantive Focus: Gender and Age Differences in FFA Factors

Gender Differences

The search for gender differences in personality research has a long history (e.g., Feingold, 1994; Hall, 1984; Maccoby & Jacklin, 1974) and gender is one of the most widely studies correlates of personality. In a meta-analysis of gender differences in FFA traits, Guo (1995) reported that women have substantially higher levels of Agreeableness and Neuroticism than men, but that gender differences were small for the other FFA factors. For Dutch adolescents, Klimstra, Hale, Raaijmakers, Branje and Meeus (2009) found that girls had consistently higher scores than boys for Neuroticism (lower Emotional Stability), Agreeableness, and Conscientiousness, with tendencies toward higher scores for Openness and Extraversion that varied with age and birth cohort. Donnellan and Lucas (2008) found that across a wide range of ages, women scored consistently higher than men on Neuroticism, Extraversion, Agreeableness, and, to a lesser extent, Conscientiousness, but that gender differences on Openness varied with nationality (higher scores for German women than German men, but lower scores for British women than British men). These gender differences did not vary substantially with age or educational level. In a cross-cultural study in 36 countries (Costa, Terracciano & McCrae, 2001; McCrae & Costa, 2001), women typically had higher scores than men on Neuroticism and Agreeableness, but gender differences were small for Conscientiousness. However, for Openness and Extraversion, the gender differences were not consistent across sub-facets of the broad trait factors. In apparently the largest cross-cultural study, Schmitt, Realo, Voracek and Allik (2008) reported that women had higher scores for Neuroticism, Extraversion, Agreeableness, and Conscientiousness than did men across for most of the 55 countries, but that differences in Openness were small. Interestingly, gender differences tended to be larger in countries with greater economic development, education, and health.

In summary, although there is considerable study-to-study variation in observed gender differences that may be a function of age, cohort, 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.

Age Differences

Plaster hypothesis. Developmental stability and change can be characterized by many features of the data (e.g., mean-level change, test-retest or rank-order stability, ipsative stability, structural stability; see Caspi & Shiner, 2006; Lüdtke, Trautwein, & Husemann, 2009). Mean-level change, the focus of the present investigation, refers to increases or decreases in the average level of an attribute in a population as a function of age on the basis of cross-sectional designs, longitudinal designs, or a combination of both. Caspi, Roberts, and Shiner (2005; also see Srivastava, John, Gosling & Porter, 2003) contrasted two conflicting theoretical perspectives about mean-level changes in FFA traits. On the one hand, many FFA proponents (e.g., Costa & McCrae, 1997) have argued that there is little mean-level change after reaching adulthood. Following from the widely cited passage from William James (1890/1963) suggesting that personality becomes "set like plaster" by age 30 (Costa & McCrae, 1994), Srivastava et al. (2003) referred to this as the *plaster hypothesis*. In one of the strongest statements of this theoretical perspective, Costa, McCrae, and Siegler (1999, p. 130) claimed that:

> *Despite wide differences in measures, subjects, and periods of the life span studied, all these studies concurred in finding relatively little change in the average level of personality traits and surprisingly high stability of individual differences. Barring interventions or catastrophic events, personality traits appear to be essentially fixed after age 30.*"

Alternatively, life-span developmentalists (e.g., Helson, Kwan, John, & Jones, 2002; Helson & Moane, 1987) argue that mean-level changes often occur in adulthood and that these are related to major life changes and role transitions. Caspi et al. (2005) reported that there was more change in FFA traits in young adulthood than in adolescence, and also noted that for some FFA traits there were systematic changes well past early adulthood, leading them to favor a life-span developmental perspective on FFA change. However, they found no clear evidence that these age-effects varied with gender, an issue of relevance to the interpretation of age effects of particular importance to the present investigation.

Proponents of the plaster hypothesis suggest that results like those summarized by Caspi et al. (2005) are largely consistent with predictions in that "From age 18 to age 30 there are declines in Neuroticism, Extraversion, and Openness, and increases in Agreeableness and Conscientiousness; after age 30 the same trends are found, although the rate of change seems to decrease" (McCrae et al., 2000, p. 183). Srivastava et al. (2003) referred to this as a "*soft plaster*" hypothesis and contrasted it with the original "*hard plaster*" hypothesis.

Based on a very large database of adult responses (for ages 21 to 60), Srivastava et al. (2003) found no support for the hard plaster hypothesis on any of the FFA factors, and found that support for the soft plaster hypothesis was limited to Conscientiousness. In a meta-analysis of longitudinal studies across the entire lifespan, Roberts, Walton and Viechtbauer (2006a, b) reported increases with age for Conscientiousness, Emotional Stability, Social Dominance (one facet of Extraversion), especially between the ages of 20 to 40. Agreeableness showed a steady increase over the lifespan, but particularly in old age. Social vitality (a second facet of Extraversion) and Openness increased during adolescence, and then decreased during old age. On the basis of these results, Roberts et al. (2006a) argued that their "meta-analysis clearly contradicts the notion that there is a specific age at which personality traits stop changing, as we found evidence for change in middle and old age for four of the six trait categories studied" (p. 14).

Maturity principle. Based on their review, Caspi et al. (2005) coined the term ***maturity principle***, saying, "Most people become more dominant, agreeable, conscientious, and emotionally stable over the course of their lives. These changes point to increasing psychological maturity over development, from adolescence to middle age" (Caspi, et al., p. 470). Noting that openness tends to decrease after young adulthood, they suggested that this pattern of maturity was more consistent with a capacity to become a productive member of society than a humanistic perspective of self-actualization. However, in their meta-analysis of longitudinal studies, Roberts, et al. (2006b) did find systematic increases in Openness during adolescence and no decline until old age. Nevertheless, only eight of the 92 reviewed studies included samples aged over 60 years old.

In a test of the maturity principle for longitudinal responses by two cohorts of Dutch adolescents followed for 5 years, Klimstra et al. (2009) reported some increases in all FFA factors, but particularly Agreeableness and Conscientiousness. However, they also reported some non-linearity for most of the age differences as well as some inconsistency across different cohorts. Donnellan and Lucas (2008) found that that across a wide range of ages (16 to 86 years old), scores decreased with age for Extraversion and Openness and increased with age for Agreeableness and Conscientiousness. Although there was some non-linearity in most of the factors, it was particularly marked for Conscientiousness (with the highest scores for middle-aged participants aged 40-50). For Neuroticism, the age differences varied somewhat with country.

In a methodologically sophisticated study, Allemand, Zimprich and Hertzog (2007) compared results for middle-aged (42-46 years old) and older (60-64 years old) adults at each of two time points separated by four years. Although there were age-related differences for all FFA traits except Conscientiousness, the results were not entirely consistent over cohort and longitudinal comparisons, and the effect sizes were modest. In another study based on a large and representative sample of Dutch adults ages between 16 and 91, Allemand et al. (2008) failed to replicate these findings; they observed age-related increases in agreeableness and conscientiousness, but non significant or non-systematic differences in the other traits. Unfortunately, none of these studies systematically investigated the possible non-linearity of the effects. In their large study of FFA factors for adults aged 21 to 60, Srivastava et al. (2003) found increases with age for Conscientiousness and Agreeableness, small decreases for Neuroticism and Openness, and no differences for Extraversion. Although there were some non-linear effects of age and age-by-gender interactions, these were mostly very small. Robins, Fraley, Roberts and Trzesniewski (2001) evaluated FFA factors for 18 and 19 year-olds at the start of university and then again four years later. They found a moderate decrease in Neuroticism, small to medium increases in Agreeableness, Conscientiousness, and Openness, and almost no mean-level changes in Extraversion.

Finally, in probably the most comprehensive study of age-related differences in FFA traits to date, Terracciano, McCrae, Brant and Costa (2005) explored cross-sectional and longitudinal age-related differences (covering and age span of 20 to 100 years old) using 1944 participants from the Baltimore Longitudinal Study of Aging. The results from the longitudinal and cross sectional analyses converged in showing: (a) non linear decreases in Neuroticism that tended to flatten out in old age; (b) non linear decreases in Extraversion that tended to accelerate after 60; (c) linear decreases in Openness; (d) linear increases in Agreeableness; (e) a curvilinear (inverted U shape) pattern in Conscientiousness characterized by initial increases up to age 60, followed by subsequent decreases.

In summary, although there is considerable study-to-study variation in observed age differences that may be a function of cohort, nationality, study design, age range and the particular instrument considered, there is clear support that over the lifespan, from adolescence to old-age, people become more agreeable and emotionally stable. Although results are mixed for Openness and Conscientiousness, there is some support for increases during adolescence and early adulthood, followed, perhaps, by decreases in old age. For extraversion, there are no clear results and the differences may vary for particular facets of this factor. Although these changes clearly contradict the plaster hypothesis, neither do they provide consistent support for the maturity principle. The maturity principle suggests that as individuals grow older their personalities evolve so that they become more mature, productive contributors to society. However, it is not clear whether the maturity principle applies only to changes during late-adolescence and early-adulthood or whether it also applies the middle- and late-adulthood. What appears particularly unclear is whether, as suggested by Caspi et al. (2005) increases in dominance, that were not clearly replicated in many studies, could really be taken to reflect maturity. Altough this may make sense since Caspi et al. appear to equate maturity with productivity, they do not specify what "productive maturity" means in old age, following retirement. Indeed, although very few of the previous studies included participants over 60 years of age, these studies suggest that additional changes seem to occur following this age, in apparent contradiction with the maturity principle. Observations such as these led Roberts et al. (2006b, p. 31) to conclude: "Moreover, accepting the fact that personality traits change in adulthood highlights the inadequacies of almost all theoretical positions found in personality psychology and personality development".

La Dolce Vita Effect: FFA changes in old age. Following from the Roberts et al. (2006b) critique, apparently a better characterization of age effects relevant to old age is needed. Marsh, Martin and Jackson (2010) offered an alternative perspective on aging based on multiple dimensions of physical self-concept for late-adolescents (aged 16-19; *M* age = 17 years) and older adults (aged 52-93; *M* = 63 years). Their Physical Self Description Questionnaire was designed to measure 9 specific physical factors (Health; Coordination; Activity; Body Fat; Sport; Appearance; Strength; Flexibility; Endurance) and two global factors (Global Physical; Global Esteem). Factor analyses demonstrated a well-defined factor structure that was invariant over gender and age. Age differences, not surprisingly, showed that the older adults had worse scores on all the nine specific physical factors (particularly Sport, Endurance, Health, and Body Fat). Interestingly, however, the older adults had Global Physical self-concepts that were as good as or slightly better than the adolescent age group, and significantly higher levels of global self-esteem. The authors speculated that as people grow older, their physical attributes decline and they are generally aware of this, as reflected in lower ratings on the nine specific factors. However, they also become more accepting of these effects and develop strategies to protect their sense of self that leads to positive and resilient self-esteem (e.g., Alaphilippe, 2008; Brandtstädter & Greve, 1994; Carstensen & Freund, 1994). Furthermore, self-concept is highly dependent on frame of reference effects as well as other standards. Indeed, specific physical self-concept factors are closely tied to actual performances so that they are strongly influenced by declines in these objective external standards, and they show some decline with age. However, for global self-esteem and, to a lesser extent, the global physical self-concept scale, respondents have a lot more flexibility in operationalizing the frame of reference—using social comparison processes such as comparisons with others of a similar age. This suggests that these older participants understand that they have diminished attributes in many physical areas, but apparently have come to terms with these differences in how they think about themselves globally; they become more content with themselves even though physical attributes are declining.

These heuristic speculations about the juxtaposition between age, global self-esteem, and specific components of the physical self-concept may also have relevance to changes FFA factors for older adults. In particular, existing research with older adults does not support hard or soft plaster hypotheses and support for the maturity effect is limited largely to late-adolescence and early-adulthood. Thus, for example, the Roberts et al. (2006a, b) meta-analyses, as well as the Terracciano et al. (2005) primary study, showed that Conscientiousness, Openness, Neuroticism and Extraversion decreased during old age, whilst agreeableness increased substantially. Similarly, Donellan and Lucas (2008) results suggested that the initially increasing levels of Conscientiousness may in fact start to decrease following the age of 50. Consisent with these observed differences in FFA factors and suggestions from the Marsh et al. self-concept study, individuals appear to become more self-content in old age – what we here refer to as the ***La Dolce Vita*** effect.

In Italy, the expression "*La Dolce Vita"* is used to describe the soft, slow, enjoyable, happy and self-indulgent traditional Italian way of life. Literally, *La Dolce Vita* thus means "the sweet life". Interestingly, *Dolce* also means the dessert, which is relevant to the present proposition since the dessert is the last, and often happiest or at least sweetest, part of the meal. In Italy, one way that La Dolce Vita manifests itself in old age is through the increased attachment of seniors citizens to their own city or village, where they are content to spend long afternoons in the shade, talking with long time friends, without ever feeling the need to visit neighbors from adjacent cities or counties. This interpretation is consistent with the observed results from FFA research showing that people become more agreeable and emotionally stable with age, but also become more laid back, satisfied with themselves and what they have and thus seems to feel less the need to reach out for more – less socially outgoing and extraverted, more introverted as well as less conscientious – perhaps because as people start to enjoy life, they also become less preoccupied with productivity.

Support for the La Dolce Vita effect also comes from research showing that older people report fewer negative interpersonal interactions than younger people (i.e., are more agreeable) and that when they do, they also report less negative affect (e.g. Almeida, 2005; Birditt & Fingerman, 2005; Lefkowitz & Fingerman, 2003). We also note that this La Dolce Vita effect is apparently consistent with emerging research showing that most forms of mood, anxiety, behavioural, substance abuse and personality disorders tend to decrease past 50 or 60 years old, and that onset of these problems in old age is rare (e.g. Degenhardt et al., 2008; Grant et al., 2004; Huang et al., 2009; Jackson, & Burgess, 2000; Kessler, Berglund, Demler, Jin, & Walters, 2005; Kessler et al., 2007; Lenzenweger, Lane, Loranger, & Kessler, 2007; The ESEMeD⁄MHEDEA-2000 Investigators, 2004a, 2004b). Because there is not a lot of methodologically rigorous research comparing developmental changes in FFA factors across the entire adolescent-to-late-adult age span – and particularly old age – this is a specific focus of the present investigation.

Taxonomy of measurement invariance models: Implications for applied research.

In psychological research, comparisons of group means (and even relations between variables) are based on typically implicit, untested assumptions about measurement invariance. A particularly important application of CFA techniques has been to test the assumptions about the invariance of the FFA 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 really do reflect the same construct and the measurements themselves are operating in the same way (across groups, over age and time, or across different levels of continuous variables), mean differences and other comparisons are likely to be invalid. Important issues for applied research are the implications for failures of these tests of invariance—in relation to the development of measurement instruments and the interpretation of results based on well-established measures. Although these concerns are known to many developmental researchers, they are frequently ignored in applied research. In FFA research there are few studies that address these issues and they apparently are not well-understood by applied researchers in this field.

Taxonomy of invariance. Marsh et al. (2009) introduced a taxonomy of 13 ESEM models (see Table 1) designed to test measurement invariance. Within the ESEM framework, the applied developmental and personality researcher has access to typical parameter estimates, standard errors, goodness of fit statistics, and statistical advances normally associated with CFA/SEMs (see Asparouhov & Muthén, 2009; Marsh et al., 2009). Importantly, ESEM allows applied FFA researchers to pursue appropriate tests of measurement invariance when CFA models are not appropriate. This taxonomy of invariance tests (Table 1) integrates factor analysis (e.g., Jöreskog & Sörbom, 1988; Marsh, 1994; 2007; Marsh & Grayson, 1994) and measurement invariance (e.g., Meredith, 1964; 1993; Meredith & Teresi, 2006) traditions to evaluate full measurement invariance: Configural invariance (all parameters are freely estimated in all groups; Model 1 in Table 1); *Weak measurement invariance* (factor loadings are invariant; Model 2), *Strong measurement invariance* (invariance factor loadings and item intercepts; Model 5), *Strict measurement invariance* (invariance of factor loadings, item intercepts, and item uniquenesses; Model 7). This taxonomy expands this measurement invariance tradition to include tests of latent means invariance and of the factor variance-covariance matrix and various combinations of invariance constraints across different sets of model parameters (see the remaining models in Table 1). Although these tests require full invariance of all parameter estimates for all groups, Byrne, Shavelson and Muthén (1989) argued for the usefulness of a less demanding test of partial invariance in which a subset of parameters are not constrained to be invariant.

Consequences of a lack for Invariance. Tests of the invariance of factor loadings (*Weak measurement invariance,* Model 2) are particularly important both in terms of relating FFA factors to other constructs for different groups with cross-sectional data or for evaluating patterns of relations among variables in the same group over time with longitudinal data. Indeed, all models except the configural invariance model (Model 1) assume the invariance of factor loadings. Unless the factor loadings are reasonably invariant over occasions or groups, then any comparisons must be considered suspect as the constructs themselves differ (i.e., the apples and oranges problem). However, if there is a sufficient number of items, tests of partial invariance might be warranted such that invariance of factor loadings is supported for almost all the items for each factor. Such tests of invariance might also be on basis of selecting items to be retained in early stages of instrument development.

If applied researchers want to compare latent mean differences across groups or over time, then tests of item intercept invariance (*Strong measurement invariance,* Model 5) are critical in addition to factor loading invariance. For example, assume for six items designed to measure a particular trait, three clearly favor women and three clearly favor men. These results provide no basis for evaluating gender differences in the trait in that even the direction of gender differences would depend on the particular items used to measure the trait. Furthermore, because these 6 items are only a small sample of items that could be used to evaluate this trait, the results provide only a weak basis for knowing what would happen if a larger, more diverse sample of items was sampled. Support for the invariance of item intercepts would mean that gender differences based on each of the items considered separately is reasonably consistent in terms of magnitude as well as direction. These results would provide a stronger basis of support for the generalizability of the interpretation of the observed gender differences. Although issues of non-invariance of item intercepts and differential item functioning are well known and evident in some FFA measures (e.g., Costa, et al., 2001; Marsh, Lüdtke et al., 2010), these issues have been largely ignored in FFA research (but see Marsh, Lüdtke, et al., 2010; Nye, et al., 2008; Reise, et al., 2001; Jackson et al., 2009) – due in large part to the apparent inappropriateness of CFA to FFA research (Marsh, Lüdtke, et al., 2010). In summary, a lack of invariance of item intercepts would mean that the observed group differences are not consistent across even the items used to represent a latent factor on a particular instrument and provide no basis for the generalizability of the results across a wider and more diverse set of items that could be used to represent the trait.

In order to compare FFA (manifest) scale scores (or factor scores), then the invariance of items' uniquenesses also represents an important prerequisite (*Strict measurement invariance,* Model 7). Indeed, the presence of differences in reliability (as represented or absorbed in the item uniquenesses) across the multiple groups could distort mean differences on the observed scores. However, for comparisons based on latent constructs that are corrected for measurement error, the valid comparison of latent means only requires support for strong measurement invariance and not the additional assumption of the invariance of measurement error. Hence, comparison of group mean differences based on latent-variable models like those considered here makes fewer assumptions than those based on manifest scores.

A lack of invariance in relations between factors does not compromise comparisons of latent mean differences over time or groups. However, particularly for multifactorial constructs like the FFA factors, the pattern of relations among factors might have important practical or theoretical implications. Furthermore, interpretations are likely to be complicated by heterogeneity of relations between FFA factor and other variables. In more complex models, the invariance of other parameter estimates (e.g., correlated uniquenesses or path coefficients) may also be relevant as a test of the generalizability of the results.

Issues of invariance have a long history in the development and application of standardized achievement tests in educational settings. Here issues such as differential item functioning – a lack of invariance – are evaluated routinely and even mandated by legal concerns (i.e., that tests are equally predictive for different groups). Although these issues are not considered so widely in the measurement of psychological constructs such as FFA factors, increasing methodological sophistication and the availability of appropriate statistical tools means that these approaches are likely to become more widely used. However, standards of best practice are still evolving – particularly in relation to what constitutes acceptable levels of invariance and partial invariance. An interesting perspective for applied research might be to evaluate how robust key parameter estimates and interpretations are to invariance assumptions. Thus, for example, if critical interpretations are similar for fully- and partially-invariant models, then applied researchers can have confidence in the appropriateness of the conclusions. However, if the interpretations change fundamentally for fully- and partially-invariant models, then interpretations should be made with appropriate caution. Thus, as in all applied research, the researcher has an obligation to interrogate the appropriateness of conclusions.

The Present Investigation: A Substantive-Methodological Synergy

Despite substantial research on how FFA manifest means (e.g., scale scores or factor scores) are related to age, gender and their interaction, these FFA developmental studies are typically methodologically weak. In particular, they often rely on the interpretation of mean differences based on manifest scale scores rather than latent variable models allowing correction for potentially complex structures of measurement error and the evaluation of measurement invariance assumptions implicit in such comparisons. Here we demonstrate ESEM– an integration of EFA, CFA and SEM that has the potential to overcome many of overly restrictive assumptions of CFA (that have led some to reject CFA as appropriate for FFA research) and limitations of EFA. In one of the first applications of ESEM to FFA research based on responses by late-adolescents to the 60-item NEO-FFI, Marsh, Lüdtke et al. (2010; but also see Dolan, et al., 2009 for similar conclusions) found: (a) ESEM fit the data better and resulted in substantially more differentiated (less correlated) factors than CFA; (b) support for full measurement invariance of the ESEM factor structure over gender, showing that women score higher on all NEO big-five factors; (c) support for measurement invariance over two years and for the maturity principle in late adolescence (decreases in Neuroticism, increases in Agreeableness, Openness and Conscientiousness). Based on ESEM, they addressed substantively important questions with broad applicability to psychological research that could not be appropriately addressed with traditional approaches to either EFA or CFA. In the present investigation we expand this application of ESEM to the methodologically demanding task of comparing FFA factors across the entire adolescent to old age span (ages 15-99), as well as extending ESEM in a number of ways that have important substantive and methodological implications.

Our study is based on a very large, nationally representative, cross-sectional sample (n=14,021) that covers the entire late-adolescent to very old lifespan (ages 15-99). Methodologically, we begin by comparing CFA and ESEM factor structures to test the prediction that ESEM results in a better fit and smaller correlations among FFA factors than CFA. We extend the ESEM model to test the full measurement invariance of the FFA factors over gender (based on the 13-model taxonomy presented in Table 1), evaluate a descriptive model of linear and non-linear age effects on latent ESEM factors with a multiple indicators multiple causes (MIMIC) model, and then combine the MIMIC and gender invariance models to test the invariance of age effects over gender. Next, we form six groups – representing all combinations of two genders and three (young, middle, and old) age groups—and test measurement invariance across these six groups. In evaluating this multi-group invariance model we introduce an ESEM-Within-CFA strategy that greatly enhances the flexibility of ESEM and allows us to partition latent mean differences into tests of age (linear and non-linear), gender, and interaction effects. Finally, we extend the MIMIC/multiple-group hybrid approach, by adding MIMIC age effects (linear and quadratic) to the gender-age multiple group models. In this way, we estimate the combined effects of age – based on continuous age (MIMIC) and multiple age groups – and their interaction with gender.

Substantively, and consistent with previous research, we expect women to score higher than men on Neuroticism and Agreeableness, and, perhaps, Conscientiousness and Extraversion. However, we have no clear basis for predicting how gender differences vary across such a wide age span. Based on the productive-maturity principle, we expect that particularly during the late adolescent and early adult years there will be decreases in Neuroticism and increases in Conscientiousness and Agreeableness, but we anticipate that these differences will not extend into old age. Based on the La Dolce Vita effect in old age, we expect our oldest participants to be more self-content, self-centered, less preoccupied with productivity, more laid back and happier as represented by decreases in Openness, Extraversion, Neuroticism and Conscientiousness, but increases in Agreeableness. As a substantive-methodological synergy, we address these substantively important questions with new, evolving, and apparently stronger methodology than previous FFA research, and demonstrate its broad applicability to developmental and psychological research more generally. Finally, we conclude with a discussion of limitations of the present investigation, including reliance on a cross-sectional design (and resulting caveats in relation to interpretations), personality assessment based on FFA self-report measures, and complications when responses are not fully invariant over covariates.

Methods

Sample and Materials

In the nationally representative British Household Panel Study (BHPS), households were selected using a multistage probability design in which all household members aged 16 and older were asked to participate. In wave 15 of the BHPS that is the basis of our study, FFA measures were administered in a self-completion format in late 2005 and 2006. The participants (n= 14,021, 54% women; age: 15 to 99, M = 47, SD = 19) completed the 15-item FFA instrument (Taylor, Brice, Buck & Prentice-Lane, 2009; also see Donnellan, Oswald, Baird, & Lucas, 2006, John & Srivastava, 1999; Rammstedt & John, 2007) in which three items were used to infer each factor using a 7-point scale (“1 does not apply” to “7 applies perfectly”). For more details, see the BHPS Technical Manual (Taylor, et al., 2009) and the online supplements.

Consistent with the brevity of the scales, coefficient alpha reliabilities for the FFA factors based on these data were: .67 (Neuroticism), .68 (Openness to Experience), .54 (Extraversion), .53 (Agreeableness), and .53 (Conscientiousness). However, reliability varies in part with the number of items and FFA instruments typically have much longer scales that are only moderately reliable (e.g., Costa & McCrae, 1997; Marsh, Lüdtke et al., 2010). In addition, the few retained items are intended to maximally cover broad constructs in line with original FFA-approach; this also might lead to lowered internal consistency in combination with the small number of items, but is necessary to capture the relatively broad FFA-domains. In fact, the psychometric properties of this short form and its abilities to adequately cover broad FFA constructs has been well documented in previous research (e.g., Taylor et al, 2009; also see Donnellan et al, 2006, John & Srivastava, 1999; Rammstedt & John, 2007). If the items were narrower in content, alpha might be higher, but the content of the FFA domains would not be covered as well by the short form. We also note that according to the Spearman-Brown prophecy formula, the reliability estimates for the full NEO-FFI instrument would be of a similar size if based on only three items. Thus, for example, when the reliability of a 3-item test is .5, the estimated reliability of an equivalent 12-item test (the number of items on the NEO-FFI) is .80. Reliability estimates for the 12-item NEO scales typically vary from the mid .70s to the mid .80s (e.g., Costa & McCrae, 1997; Marsh, Lüdtke et al., 2010). Hence, reliabilities observed here are reasonable in relation to other research after taking into consideration the number of items per scale. Of course, given the modest reliability, it is important to base conclusions on latent-variable models that correct for unreliability.

For all but Openness, there were two positively worded items and one negatively worded item in each factor (all Openness items were positively worded). In each case, the negatively worded item had the lowest item-total correlation (although all items-total correlations were positive following inversion of these items) and for the Agreeableness, Conscientiousness, and Extraversion factors, the elimination of the negatively worded item would have resulted in a slightly higher estimate of reliability. Consistent with these observations, preliminary results suggested a response biases associated with negatively worded items that is common in self-report instruments (e.g., Bagozzi, 1993; Corwyn, 2000; Marsh, 1986, 1996).

Statistical Analyses.

All analyses in the present investigation were conducted with Mplus 5.2 (Muthén & Muthén, 2008). The main focus is on the application of ESEM to responses to the 15 FFA items. 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 non-normality and non-independence of observations (Muthén & Muthén, 2008). The ESEM approach differs from the typical CFA approach in that all factor loadings are estimated, subject to constraints necessary for identification (for further details, see Asparouhov & Muthén, 2009; Marsh et al., 2009). Although there are many methodological and strategic advantages to ICM-CFAs, these models typically do not provide an acceptable fit to the data. In related research, Marsh (2007; Marsh, Hau, & Grayson, 2005) argued that few multidimensional assessment instruments met even minimal standards of goodness of fit based on CFA. Part of the problem, we argue, is undue reliance on overly restrictive independent cluster models of confirmatory factor analysis (ICM-CFA) in which each item is hypothesized to load on one and only one factor. This failure to achieve acceptable levels of fit has led to many compensatory strategies that are dubious, counterproductive, misleading, or simply wrong (e.g., analysis of item parcels). Furthermore, the misspecification of factor loadings (constraining them to be zero when they are not) usually leads to distorted factors with over-estimated factor correlations that might lead to biased estimates in structural equation models (SEMs) incorporating other outcome variables (Asparouhov & Muthén, 2009; Marsh et al., 2009a, 2009b; Schmitt & Sass, 2011). Indeed, even when CFA does provide an acceptable fit to the data, ESEM not only provides a better fit but also results in latent factors that are much more differentiated (i.e., less correlated). This is not surprising in that ESEM uses two estimates of overlap between factors (overlap in factor loadings and correlation between factors), whereas CFA uses one estimate (correlation between factors).

Following Marsh, Lüdtke, et al. (2010) we used an oblique geomin rotation (the default in Mplus) with an epsilon value of .5. There were few missing responses (less than 1%), that were handled with the full-information MLR estimator to correct for missing data. Because of the design of the BHPS in which respondents are clustered within households, we used the Mplus complex survey design option to control the clustered design and adjust standard errors. Sampling weights were also taken into account in the analyses.

In general, the use of ex-post facto correlated uniquenesses (CUs) should be avoided (e.g., Marsh, 2007), but there are some circumstances in which a priori CUs should be included (Jöreskog, 1979); Marsh & Hau, 1996). For self-report surveys that include a mixture of positively and negatively worded items, it is typical to find method effects associated item wording (Marsh, Scalas & Nagengast, 2010). In the present application, 4 out of 15 items (one each for four of the five factors) were negatively worded. We thus adopted a standard, a priori approach to address this potential artifact by specifying CUs relating the responses to each of these negatively worded items (e.g., Marsh, 1996). In preliminary analyses, we compare solutions with and without these CUs to evaluate the appropriateness of this strategy.

Multigroup tests of invariance and latent mean differences. Tests of invariance and latent mean differences pursued here are based on the taxonomy of invariance tests (Table 1; see earlier discussion). Multi-group tests of invariance typically consist of comparisons across only two groups or, possibly, more than two groups that represent different levels of the same variable (e.g., multiple age groups). However, the logic of this strategy is easily extended to include all combinations of groups representing two or more variables. Although mean comparisons based on such groups are typical in ANOVA studies based on manifest variables, these comparisons are also based on the assumption of strict invariance (i.e. loadings, intercepts, uniquenesses) across all groups reflecting the main effects of both variables and their interaction. These assumptions – particularly in relation to groups formed by the interaction of two or more variables, are very rarely tested, and there is little basis for knowing how robust conclusions are in relation to these untested assumptions. Although this has apparently not been previously verified in published FFA research, we demonstrate an extension of the multiple-group ESEM model to test invariance across 6 groups representing all combinations of the two genders and three age categories: 15-30 (N = 3194, M = 22.5; SD = 4.5), 31-60 (N = 7211, M = 45.1; SD = 8.6), 61-99 (N = 3678, M = 72.1; SD = 7.8). These categories correspond to roughly late-adolescent/young adulthood, middle age, and older age categories and have been considered in previous research. Thus, for example, 30 is the age at which Costa and McCrae (1994) proposed that personality “becomes set like plaster” whilst 60+ is the upper age category considered by a number of previous studies (e.g., Terracciano, et al., 2005; Allemand, et al., 2007) and age at which Roberts et al. (2006b) noted that there was a dearth of research. Although such analyses have the obvious limitation that the continuous age variable is divided into broad categories with a potentially serious loss of information. Here we introduce a MIMIC/multigroup hybrid model to address this problem.

MIMIC/Multiple-group Hybrid Model of Age Effects (see Supplemental Materials for further discussion). For studies of age differences in FFA factors, the tests of invariance become even more complex in that age is a continuous variable rather that a natural categorical variable with a few discrete groups (like gender). There are traditionally two approaches to this problem. The MIMIC model regresses the latent variables (the FFA factors) onto other variables (continuous, like age, or categorical, like gender). However, only the invariance of factor means and item intercepts (by the addition of direct effects between the covariate and the items) can be tested. In the multiple group approach, it is possible to pursue the more rigorous tests of invariance presented in Table 1. However, for continuous variables, these tests require researchers to transform continuous variables into a relatively small number of categories that constitute the multiple groups. Marsh, Tracey and Craven (2006) proposed a hybrid approach involving an integration of interpretations based on both MIMIC and multiple group approaches. Here we extend this approach in several ways: demonstrating how the MIMIC and multiple group approaches can both be incorporated into a single ESEM model, adding the MIMIC age (linear and quadratic effects) variables to the multiple group model (based on sex-age groups). This allows us to evaluate more formally whether information in the continuous age effects are lost by forming age categories and, if so, to estimate the combined age effects due to both operationalisations of age (continuous and categorical). The interaction between gender and age is substantively important to interpretations of both gender and age effects. Although tests of invariance have rarely been applied to the interaction of two variables, we illustrate how the use of ESEM to the hybrid integration of multiple-group and MIMIC models can be extended to include interactions between variables.

ESEM-Within-CFA Model (see Supplemental Materials for further discussion). Despite the flexibility of the ESEM approach, we note that there are some aspects and extensions of traditional SEM models that cannot readily be implemented with ESEM as currently operationalized in Mplus (e.g., constraints on group specific correlations among factors, partial invariance of factor loadings, tests of higher-order factor models, latent curve models based on multiple manifest indicators of the longitudinal construct, partially invariant factor mixture models, etc. also see Asparouhov & Muthén, 2009; Marsh, Lüdtke et al., 2010; Marsh et al., 2009). Of particular relevance to the present investigation, applied researchers cannot easily place constraints on latent means estimated in multiple group models to test linear and non-linear effects based on a single grouping variable (e.g., age) or the interaction between two grouping variables (e.g., age-by-gender interactions). Here we propose an extension of the ESEM approach to address this limitation – what we refer to as ESEM-Within-CFA Models. Although not a major focus of the present investigation, this ESEM-Within-CFA strategy can easily be applied to many other situations in which CFA models cannot be evaluated with ESEM, thus further enhancing the flexibility of ESEM.

Goodness of fit. CFA/SEM research typically focus on the ability of a priori models to fit the data as summarized by sample size independent fit indexes (e.g., Marsh, 2007; Marsh, Balla, & McDonald, 1988; Marsh, Balla & Hau, 1996; 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 $\chi 2$ test statistic and evaluation of parameter estimates. For the TLI and CFI values greater than .90 and .95 are typically interpreted to reflect acceptable and excellent fit to the data. For the RMSEA values of less than .05 and .08 are typically interpreted to reflect a close fit and a reasonable fit to the data, respectively (Marsh, Hau, & Wen, 2004). However, we emphasize that these cut-off values only constitute rough guidelines (Marsh, 2007; Marsh et al., 2005; also see Marsh, Hau, Balla & Grayson, 1998). Furthermore, because there are very few applications of ESEM – and none that fully evaluate the appropriateness of the traditional CFA indexes of fit – the relevance of these CFA indexes and the proposed cut-off values are not clear (Marsh et al., 2009).

It is typically more useful to compare the relative fit of different models in a nested or partially nested taxonomy of models designed a priori to evaluate particular aspects of interest than 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 $\chi 2$-test that led to the development of fit indexes (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 requires a change in CFI of less than .01 (Cheung & Rensvold, 2002; Chen, 2007) or a change in RMSEA of less than .015 (Chen, 2007). Marsh (2007) noted that some indexes (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 (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 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.

RESULTS

FFA Factor Structure: ESEM vs. CFA

The starting point for the present investigation is to test our a priori hypothesis that the ESEM model provides a better fit to FFA responses than a traditional CFA model in which items are constrained to have zero factor loadings on all factors but the one each was designed to measure (hereafter referred to as the independent clusters model, or ICM-CFA). Indeed, as emphasized by Marsh et al. (2009), the ESEM analysis is predicated on the assumption that ESEM performs noticeably better than the ICM-CFA model in terms of goodness of fit (Table 2) and construct validity of the interpretation of the factor structure.

In our study, the ICM-CFA solution does not provide an acceptable fit to the data (TLI = .687; CFI = .761; RMSEA = .076; see TGCFA1A in Table 2). The next model incorporates a priori CUs (to control for method effects associated with negatively worded item; see earlier discussion); results are still inadequate, albeit improved (TLI = .722; CFI = .804; RMSEA = .072; see TGCFA1B in Table 2). Apparently, all existing standards of acceptable fit would lead to the rejection of the ICM-CFA model. The corresponding ESEM solutions fit the data much better. Although the fit of the model with no a priori CUs is marginal (TGESEM1A: TLI = .889; CFI = .958; RMSEA = .045), the inclusion of CUs results in a much better fit to the data (TGESEM2A: TLI = .948; CFI = .983; RMSEA = .031).

It is also instructive to compare parameter estimates based on the ICM-CFA and ESEM solutions (Table 3). In both models, the main factor loadings tend to be modest, with few loadings greater than .8 and some factor loadings less than .50. Although CFA factor loadings (M = .60, Md = .63) are somewhat higher than for the ESEM model (M= .57, Md = .56), the differences are typically very small and the pattern of factor loadings is very similar for the CFA and ESEM solutions. However, the R-Square estimates of communalities are slightly higher for the ESEM solution (M = .44, Md = .43) than for the CFA solution (M = .40, Md = .40). Again, however, the pattern of results is highly similar. We also note that the factor loadings associated with the negatively worded items are consistently lower than those of the positively worded items (the same pattern was evident in the unreported models without CUs, consistent with preliminary analyses of coefficient alpha estimates).

A detailed evaluation of the factor correlations among the FFA factors demonstrates a critical advantage of the ESEM approach over the ICM-CFA approach. Although patterns of correlations are very similar, the CFA factor correlations (-.17 to +.68; M absolute value = .33, Md absolute value = .38) are larger than the ESEM factor correlations (-.07 to +.41; M absolute value = .16, Md absolute value = .16). Thus, for example, the correlation between Agreeableness and Conscientiousness is +.40 for the CFA, but only +.10 for the ESEM. In this respect, the ESEM solution is more consistent with a priori predictions (see McCrae et al., 1996) that CFA factor correlations are positively biased by the failure to include cross-loading as in the ESEM solution.

In summary, the ESEM solution is clearly superior to the CFA solution, both in terms of fit and in the distinctiveness of the factors that is consistent with predictions based on the FFA. 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 much better than what has typically been achieved in previous attempts to analyse the FFA through CFAs conducted at the item level. Because the fit of the CFA model is so bad, it would be inappropriate to pursue analyses based on this model and even more dubious to base analyses on manifest scale scores computed on the implicit assumption that the fit of this CFA model is acceptable. Hence, this is clearly a demonstration of why ESEM might be so important to FFA research, as well as psychological and social science research more generally.

Invariance over Gender.

How similar is the FFA structure for men and women? Are there systematic gender differences in latent means, and are the underlying assumptions met to justify interpretations of these results? To address these questions, we applied the taxonomy of 13 ESEM models described earlier (see Table 1). 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 based on the total sample indicate that a priori CUs are necessary to achieve even a good 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.

The two-group model with no invariance constraints (MG1A in Table 4) provides a marginal fit to the data (TLI=.889, CFI=.958; Table 4). However, consistent with earlier results, the inclusion of the set of a priori CUs substantially improves the fit (TLI=.942, CFI=.981; MG1B). Importantly, constraining these a priori CUs to be invariant over gender (MG1C in Table 4) resulted in almost no change in fit. For fit indexes that control for parsimony, the fit is essentially unchanged or slightly better for MG1C than MG1B (.942 to .944 for TLI; .033 to .032 for RMSEA). For the CFI that is monotonic with parsimony, the change (.981 to .980) is clearly less than the .01 value typically used to support invariance constraints. These results are substantively important, demonstrating that the sizes of the six a priori CUs are reasonably invariant over gender. A similar pattern is also evident in MG2 (factor loadings invariant) and MG3 (factor loadings and uniquenesses invariant). The consistency of this pattern of results over the different models provides support for the inclusion of these a priori CUs. Thus, in order to facilitate communication of the results, we subsequently focus primarily on models that include invariant CUs are included (models labeled “C” in Table 4; e.g. Model MG1C for Model 1).

Weak factorial/measurement invariance tests whether the factor loadings are the same for men and women. Model MG2C (along with MG2A and MG2B) tests the invariance of factor loadings over gender. The critical comparison between the more parsimonious MG2C (with factor loadings invariant) and less parsimonious MG1C (with no factor loading invariance) supports the invariance of the factor loadings over gender: fit indexes that control for model parsimony are as good or better for the more parsimonious MG2C (TLI=.960 vs. .944; RMSEA=.028 vs. .032), whilst the difference in CFI that is monotonic with complexity is only slightly smaller (CFI=.977 vs. .980) and clearly less than the .01 difference typically used to argue for the less parsimonious model. We interpret these results to provide good support for weak measurement invariance – the invariance of factor loadings.

Strong measurement invariance requires that item intercepts – as well as factor loadings – are invariant over groups. The critical comparison is thus between Models MG2C and MG5C and tests whether differences in the 15 intercepts can be explained in terms of five latent means. The fit of MG5C (TLI= .946, CFI=.966) is reasonable, but not as good as the fit that of the corresponding MG2C (TLI= .960, CFI=.977). This suggests that gender differences at the level of items intercepts cannot be fully explained in terms of the latent means, i.e. that there is evidence of differential item functioning. Because invariance of item intercepts is so central to the evaluation of latent mean differences, we pursued alternative tests of partial invariance of item intercepts. Based on (ex post facto) modifications in which we freed parameters one at a time, we identified 4 (of 15) item intercepts that contributed most to the misfit associated with the complete invariance of item intercepts in Model MG5Cp (the additional p indicating that there is partial rather than full invariance; see footnote 1). The results support partial invariance of item intercepts. For example, fit indexes that control for parsimony are nearly the same for MG5Cp compared to MG2C (.960 vs. .960 for TLI, .027 vs. .027 for RMSEA), whilst the difference in CFIs (.975 vs. .977) is less than the .01 value that would lead to the rejection of constraints imposed in MG5Cp. However, the interpretation of these results is cautioned by the ex post-facto nature of these modifications.

Strict measurement invariance requires that item uniquenesses, item intercepts, and factor loadings are all invariant over the groups. Here, the critical comparison is between models MG5Cp and MG7Cp. Model MG7Cp does provide evidence of a good fit to the data (TLI = .959, CFI = .972, RMSEA = .028) that is similar to that of MG5Cp. Furthermore, comparisons of all the pairs of models that test the invariance of the uniquenesses (MG3C vs MG2C; MG6C vs MG4C; MG7Cp vs MG5Cp; MG9Cp vs MG8Cp; MG11Cp vs MG10Cp; MG13Cp vs MG12Cp) consistently result in a change in CFIs that are under than the .01 value typically used to support the more parsimonious model with uniquenesses invariant.

Factor variance-covariance invariance is typically not a focus of measurement invariance, but 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 FFA factors across groups is common, these are typically based on manifest scores that do not control for measurement error and makes implicit invariance assumptions that are rarely tested. Here, the most basic comparison is between Models MG2C (factor loadings invariant) and MG4C (factor loadings and factor variance-covariances invariant). The results provide reasonable support for the additional invariance constraints, both in terms of the values for the fit indexes and their comparison with MG2C. For example, fit indexes that control for parsimony are nearly the same for MG4C compared to MG2C (.959 vs. .960 for TLI, .028 vs. .027 for RMSEA), whilst the differences in CFIs (.973 vs. .977) are less than the .01 cut-off value that would lead to the rejection of constraints imposed in the more parsimonious MG4C.

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 (MG10Cp-MG13Cp in Table 3) in the taxonomy all constrain mean differences between males and females to be zero – in combination with the invariance of other parameters.

Again, several models that could be used to test for gender mean invariance: (a) MG5Cp vs. MG10Cp; (b) MG7Cp vs. MG11Cp; (c) MG8Cp vs. MG12Cp; (d) MG9Cp vs. MG13Cp. However, all these comparisons lead to the conclusion that latent means representing the FFA factors differ systematically for men and women. Based on these results, we chose Model MG7Cp (Table 4) as the best fitting model. Based on this model, latent means for women in SD units were systematically higher than for men on Agreeableness (.27), Conscientiousness (.11), Extraversion (.30), and Neuroticism (.60), but lower for Openness (-.42).

In summary, there is reasonable support for the invariance over gender of factor loadings and partial invariance item intercepts (partial strong measurement invariance) that provide a justification for the interpretation of gender differences based on latent means. The observed gender differences were consistent with a priori predications. We now extend these analyses to evaluate age differences in the FFA factors and whether gender differences vary as a function of age.

MIMIC Models of Age, Gender and their Interaction.

Do FFA factors change with age? Are these effects of age linear or non-linear? How do these age effects vary with gender? Is there support for plaster hypotheses, maturity effects for adolescent and early adult ages, and/or the La Dolce Vita effect in old age? We address these questions with a set of three MIMIC models (see Tables 2 and 5). We begin with models including only linear and quadratic components of age and then extend these to include gender and its interaction with age.

MIMIC Models of Age Effects. We begin with the ESEM model based on the total group (TGESEM1B, in Table 1) and add linear and quadratic components of age to this model. This is a standard ESEM application, combining the ESEM approach with the traditional MIMIC model. Although the MIMIC model is limited in terms of testing invariance in relation to most parameters in the factor solution – particularly factor loadings – it allows for the verification of intercept invariance.

We begin with a restrictive MIMIC model that includes the linear and quadratic effects of age on each of the FFA factors (MIMICAge1). Age is based on a continuous score and item intercepts are assumed to be completely invariant over age (no direct effects of age are specified on the FFA items). This means that linear and quadratic age effects on each indicator are fully explained by the age effects on the latent factors. The fit for this model is reasonable (MIMICAge1, Table 1; CFI=.966, TLI = .916, RMSEA = .037), but not perfect. In a second model, we used post-hoc modification indexes to evaluate partial invariance models. Based on these results, we freed the linear effects of age on three indicators. Hence, intercepts were completely invariant for two FFA factors and partially invariant for three FFA factors (i.e., one of three intercepts was freed for each of three factors). However, there was no evidence of partial invariance for the quadratic age effects. Allowing for partial invariance improved the fit of the model (MIMICAge2, Table 2; CFI=.976, TLI = .936, RMSEA = .032).

In the final MIMIC model with age effects, we constrained all of the quadratic effects of age on the FFA factors to be zero. The fit of this model was clearly worse (MIMICAge3, Table 2; CFI=.964, TLI = .912, RMSEA = .037), demonstrating that there are non-linear as well as linear relations between age and FFA factors. The detailed results from the final model (MIMICAge2) are reported in Table 5 and indicate that there are statistically significant linear age effects on all FFA factors (positive for Agreeableness; negative for Conscientiousness, Extraversion, Neuroticism, and Openness). However, there are also statistically significant non-linear effects of age on all FFA factors (U-shaped for Agreeableness and Extraversion; inverted U-shaped for Conscientiousness, Neuroticism, and Openness). We will return to these effects, when we evaluate age-by-gender interactions in the next section.

MIMIC Models of Age and Sex Effects. We next add three new effects to the previous ESEM-MIMIC models: the main effects of gender and the interactions between gender and the linear and quadratic components of age. Again, we begin with a model that assumes the full invariance of the items intercepts (i.e., no effects of any of the covariates on FFA indicators that cannot be explained in terms of FFA factors). The fit for this model is reasonable (MIMICAge*Sex1, Table 2; CFI=.966, TLI = .924, RMSEA = .032), but the inclusion of partial invariance of item intercepts (freeing four paths of the 75 paths relating the 5 covariates to the 15 FFA indicators) results in a modestly improved fit to the data (MIMICAge*Sex2, Table 2; CFI=.974, TLI = .940, RMSEA = .028). In the final MIMIC model, we constrain all of the age-by-gender interactions to be zero. This model provides a good fit to the data (MIMICAge*Sex3, Table 2; CFI=.973, TLI = .945, RMSEA = .027). Indeed, fit indexes that take into account parsimony are actually better for this model without interaction effects than the corresponding model with interaction effects. These results demonstrate that there are almost no age-by-gender interactions for these data.

The results from these models are reported in Table 5 and show that there are statistically significant gender differences for all FFA factors, with women scoring higher than men for Agreeableness, Conscientiousness, Extraversion, and Neuroticism, but lower for Openness. As gender is nearly orthogonal to age, the age effects are nearly identical to those already discussed. Graphs of these results are presented in

Figure 1 and illustrate the sizes of these effects in standard deviation units. Whilst there are clear gender differences (particularly for Neuroticism), they are not large relative to the age effects. Although there is some non-linearity in the age effects, only for Conscientiousness is there a clear maximum or minimum where the effect of age changes direction. Of particular relevance, the results show that the age effects are essentially the same for men and women.

There are, of course, potentially serious limitations of the MIMIC models. In particular, they are based on the assumption of strict measurement invariance (Model 7 in Table 1: the invariance of factor loadings, items intercepts and uniquenesses in relation to the linear and non-linear components of age, gender, and the linear and non-linear age-by-gender interactions). Whilst it is possible, as we demonstrated, to test and relax the assumption of intercept invariances, it is not possible even to test the invariance of uniquenesses and factor loadings in a MIMIC model. For the main effects of gender, we have already demonstrated that there is reasonable support for the invariance of factor loadings and uniquenesses and at least partial invariance of the intercepts (see Table 4). Even though age is a continuous variable, it is possible to construct age groups and test the set of 13 invariance models (Table 1) in relation to these groups. However, this would involve an obvious loss of information in transforming a continuous variable into discrete groups. Nevertheless, the invariance of parameter estimates in relation to the age-by-gender interaction effects is a potentially more difficult limitation to which we now turn.

Multiple Group Models of Age, Gender and their Interaction.

For the purposes of analyses in this section, we consider a multiple-group ESEM model with 6 groups representing all combinations of 3 age groups (young, middle, old) and two gender groups (male, female). Tests of invariance in relation to these six groups reflect the main and interaction effects of age (linear and quadratic) and gender. Latent means based on these groups are similar to those based on the MIMIC model already discussed, but differ in two particularly important ways. First, this multiple-group approach is much more flexible in terms of testing the strict invariance assumptions implicit (but untestable) in the MIMIC model. Second, the multiple group approach is based on age groups rather than age as a continous variable. We return to a discussion of these differences after presenting the results.

The configural invariance model provided good support for the FFA model (Model MAG1, Table 6; CFI= .979, TLI = .943, RMSEA = .034), and fit indexes that controlled model parsimony were even better when factor loadings were constrained to be equal over the six age-gender groups in Model MAG2 (TLI =

.950, RMSEA = .031). The invariance of uniquenesses for all 15 items across the six groups was not supported, but there was reasonable support for partial invariance (MAG3p: CFI = .957, TLI = .950, RMSEA = .031). This pattern of partial invariance of uniquenesses was used in all subsequent models with invariance constraints on uniquenesses. Similarly, whilst strong measurement invariance (Model 5) – complete invariance of all 15 intercepts across all 6 age-gender groups – was not supported, there was reasonable support for the partial invariance of intercepts (MAG5p: CFI = .957, TLI = .948, RMSEA = .032). Putting together these two sets of constraints, there was reasonable support for partial strict measurement invariance in relation to the complete invariance of the loadings, partial invariance of uniquenesses and partial invariance of intercepts (MAG7p: CFI = .953, TLI = .948, RMSEA = .032).

Tests of the invariance of the latent factor variance-covariance matrix, as is the case with other comparisons, could be based on any pair models in Table 6 that differ only in relation to the factor variance-covariance matrix being free or not. The most basic comparison (MAG4 vs. MAG1) suggests that support for invariance of the factor variance-covariance matrix is questionable (ΔCFI=.016, ΔTLI=.007). Other pairs of models in Table 6 that differ only in relation to the factor variance-covariance matrix being free or not also show lack of support for the invariance of the factor variance-covariance matrix over time. However, because these parameters are not central to the comparison of FFA latent means across the 6 age-gender groups, we did not pursue a strategy of partial invariance.

Finally, we are now in a position to address the invariance of the latent factor means across all six groups that is a major focus of our study. Submodels MAG10p- MAG13p each test the invariance of latent means in combination with the invariance of other parameter estimates. However, for each pair of models, the fit of the model positing no latent mean differences is systematically poorer than the corresponding model in which latent mean differences are freely estimated: differences in CFI (.077 to .082), TLI (.075 to .089) based on comparisons of submodels MAG10p vs. MAG5p, MAG11p vs. MAG7p, MAG12p vs. MAG8p, and MAG13p vs. MAG 9p. Hence, there is clear evidence that the latent means differ systematically across these six age-gender groups. Thus, we retained model MAG7p (partial strict invariance) as the final model. The results from this model are presented in the second column of figure 1 and in the left-hand section of Table 7. However, the pattern of results is nearly identical for all four profiles.

It is also relevant to evaluate the consistency of the mean differences based on this multigroup approach with earlier results based on the MIMIC model, particularly given that the two approaches are based on very different assumptions. In order to facilitate comparisons we also included the results from the preceding MIMIC-Sex*Age2 model, that included five terms (age-linear, age-quadratic, sex, sex-by- age-linear and sex-by- age-quadratic), in the first column of Figure 1 and on the right-hand section of table 7. For both models, group differences from each approach were transformed into standard deviation units so that mean differences are in terms of typical effect sizes. Graphs of the results from the multiple-group approach (see Figure 1) show essentially the same pattern of results as already presented for the MIMIC approach. Visually, these graphs demonstrate the estimated effects for both groups are very similar, adding confidence in the interpretations based on each approach. This suggests that – at least in this application – the MIMIC approach is apparently reasonably robust in relation to its implicit untestable invariance assumptions, whilst the multiple group approach is reasonably robust in relation to information lost in forming age categories from the continuous age values.

In summary, the MIMIC approach provides convenient tests of the statistical significance for each of the effects of gender and age. For the multiple-group approach, it is also possible to construct contrasts on the latent mean differences to test these effects. For ESEM models this can be done by converting the ESEM model into a CFA model (see Appendix 1 for a discussion of this ESEM-within-CFA conversion and contrasts as operationalized in Mplus). Because of the very large sample sizes, almost all these effects are statistically significant. Nevertheless, there is a reasonably good correspondence between the direction and even the relative sizes of tests based on the multiple-group approach and the those already evaluated with the MIMIC approach. We now integrate these two approaches – the multiple group model and the MIMIC model – into a single analytic framework that overcomes some of the limitations of both approaches.

A Hybrid Model of Multiple Group and MIMIC Models of Age, Gender and their Interaction.

Thus far, starting with the ESEM model, we have juxtaposed the results from the corresponding MIMIC and multiple-group models, using each to cross-validate the results of the other. Particularly when there is such good correspondence between the two, this visual comparison might be sufficient. However, we now combine the two approaches to form a hybrid model that integrates the advantages of both into a single model. We use this hybrid model to determine whether there are statistically significant and substantively meaningful differences from one that cannot be explained by the other. In order to accomplish this, we add the MIMIC effects of age (linear and quadratic) to the 6 groups (3 age groups and 2 gender groups) multiple group model MAG7p (Table 6).

We begin by providing a more meaningful reference against which to compare results for models using this hybrid (MIMIC-MAG) approach based on two preliminary models. The first (MIMIC-MAG0 in Table 8) posits that there are no MIMIC age effects (age effects are included in the model, but constrained to be zero). This provides a lower bound for subsequent models. The second (MIMIC-MAGS in Table 8) is a saturated model in which paths from linear and quadratic MIMIC-age variables to all 15 FFA indicators are freely estimated in all groups. The comparison to these two models is a critical comparison in that if the difference is small or non-significant, it means that no information was lost in using the age groups instead of the continuous age variable. Particularly the indexes that control model parsimony suggest that the difference between these two model is not substantial (TLI = .932 vs. .948; RMSEA = .033 vs. .030). This implies that the MIMIC model with continuous age variables does not contribute much beyond what can be explained by the multiple-group model with discrete age categories.

Next we explore what aspects of the MIMIC age variables were critical. In MIMIC- MAG1 we included only the effects of MIMIC L-Age and Q-Age effects on latent means which were freely estimated. In MIMIC-MAG2 we added the partial invariance of intercepts identified in previous MIMIC models (see MIMICAge2 in Table 2), and evaluated if these were invariant across the 6 age-by-gender groups in the next two models. The results from these models suggest that these effects were invariant across 5 of the six groups and had to be freed in one group (young males). In MIMIC-MAG5 model, we constrained all MIMIC quadratic age effects to be zero. Particularly in relation to indexes that control for parsimony, the fit for this model (MIMIC-MAG5; TLI=.944, RMSEA = .030) was as good or better than the corresponding model that included freely estimated MIMIC quadratic effects of age (MIMIC-MAG4; TLI=.942, RMSEA = .030).

In the final models we explored various constraints on the MIMIC linear age effects. In previous MIMIC models we noted that there were very little effects of age-by-gender interactions. In MIMIC-MAG6, we evaluated this possibility by constraining all the MIMIC age effects to be equal for men and women in the same age group. In support of this constraint, there was almost no change in the fit of the model. Next we constructed a reduced model in which the smallest effects of MIMIC age were constrained to be zero, retaining 14 of the possible 30 effects (i.e., 5 FFA factors x 6 age-by-gender groups). For 12 of these 14 effects, there were matching effects for men and women within each age group. Again, constraining the effects to be invariant across gender within each of the age groups had no effect on the fit indexes.

In summary, the systematic evaluation of the hybrid MIMIC-multiple-group models showed that the

MIMIC models did not contribute much beyond what could be explained by the multiple group models in terms of age effects. The relatively small differences were limited primarily to linear effects of age in the MIMIC models and these effects within each age group were similar for men and women. Based on Model MIMIC-MAG6, we graphed the combined effects gender, age, and their interaction based on the combined effects in the multiple-group and MIMIC models (see Figure 1). This graph differs from the graph based on the multiple groups in that for each of the six age-gender groups, the additional effects of MIMIC age are added. Clearly this is the best representation of age and gender effects in our data. However, consistent with our interpretations of the statistical models, the graph based on this extended hybrid approach shows essentially the same pattern of results as is observed in results based on the separate MIMIC and multiple group approaches also shown in Figure 1.

Discussion, Implications and Directions for Further Research.

The present investigation is a substantive-methodological synergy, applying new and evolving methodological innovations to explore an ongoing substantive issue with important theoretical and practical implications for FFA and developmental research. The result of this synergy is one of the methodologically strongest studies of how FFA factors vary with gender and age. A particular design strength of the study is the use of a nationally representative sample including a wide age range. The changes in the FFA factors with age have important substantive implications for theoretical models in FFA research. The ESEM model provides clear support for the FFA factors in relation to goodness of fit that is better than the traditional CFA model. This is an important contribution in that very few studies based on any FFA instruments have been able to achieve an acceptable level of fit starting at the level of the individual item. Whilst most previous research is based on scale scores that are a crude representation of the FFA factors, our results are based on latent ESEM factors. These ESEM models better represent the underlying FFA factors, control for measurement error, and allow us to address issues that cannot be studied with manifest scores (i.e., aggregated scale scores or factor scores).

Summary of Substantive Implications.

Sizes of Correlations Among FFA Factors. FFA factors are posited to be relatively uncorrelated, but McCrae et al. (1996) and others (e.g., Dolan et al., 2009; Marsh, Lüdtke et al., 2010) argue that the application of traditional CFA models leads to inflated correlations among the FFA factors. Our results support this contention in that correlations among FFA factors defined by CFA were systematically and substantially higher than those among the corresponding ESEM FFA factors. In general, if there are at least moderate cross-loadings in the true population model and these are constrained to be zero as in the typical CFA model, the estimated factor correlations are likely to be inflated and the differences can be substantial (Asparouhov & Muthén, 2009; Marsh et al., 2009; 2011; Schmitt & Sass, 2011). 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 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 counter-balancing biases). In all CFA applications, factor correlations will be at least somewhat biased unless all non-target 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. For example, the distinctiveness of the age and gender differences across the FFA factors depends at least in part on the distinctiveness of the underlying FFA factors and how they are represented. We also note that whatever the true correlation among the factors, the estimated correlations are likely to be inflated in ICM-CFA analyses that constrain cross-loadings to be zero and these biased estimates distort the pattern of relations between FFA factors and other variables of interest.

Plaster Hypothesis. According to the “plaster hypothesis” changes in personality end – or at least slow down substantially – after age 30 (i.e., personality is set in plaster). Consistent with a growing body of research based on manifest measures, our research clearly refutes both strong and weak versions of the plaster hypothesis in relation to mean-level changes in FFA factors. All three sets of graphs in Figure 1 show that there are consistent changes in FFA latent means across the entire late-adolescent, adult, and old age range from 15 to 99. Indeed, only one of the FFA factors (Extraversion) suggests that there is even a decline in the rate of change with age. For two FFA factors (Agreeableness and Neuroticism) the rate of change is systematically larger – not smaller – in late-adulthood. For one of the FFA factors (Conscientiousness) even the direction of change is different for older adults (there are substantial increases in late-adolescence and early adulthood, but systematic declines in middle and late adulthood). Although our study is consistent with other research leading to the rejection of the plaster hypothesis, our basis for doing so is stronger in terms of the methodology and age-range. The plaster hypothesis is a dying urban myth that should be dropped from the FFA research literature.

Maturity principle. The maturity principle suggests that as individuals grow older their personalities evolve so that they become more mature, although this hypothesis appears to equate maturity with productive contribution to society. This productive-maturity principle has typically been formulated as to implicitly reflect a constant evolution across the lifespan, something that the non linear results obtained in the present investigation, as well as in all of the preceding studies in which participants older than 60 were included, clearly refute. At least superficially, some of our results may appear consistent with the productive-maturity principle at least when this hypothesis is taken to reflect FFA development in late adolescence and early adulthood–particularly the decrease in Neuroticism, the increase in Agreeableness and some of the early changes in Conscientiousness. However, when the maturity-principle is taken to reflect lifelong development, complications emerge. Whilst there are increases in Agreeableness and decreases in Neutroticism with age, the changes tend to be larger for older adults than younger adults. Does this mean that older individuals mature more than younger ones in these factors? For Conscientiousness, the increases are limited primarily to late-adolescence and early adulthood. Starting in middle-adulthood, there is a dramatic decline in Conscientiousness. Does this mean that there is a decline in “maturity” beyond middle age or simply that alternative processes emerge? Whilst predictions based on the maturity principle have been ambiguous in terms of Extraversion (and may even differ across subfacets of this factor), the decline with age observed here is apparently inconsistent with current formulations of the maturity principle suggesting that increases in dominance (a facet of Extraversion) reflect increasing maturity, or more appropriately, productive maturity. Finally, the steady decline in Openness observed here (and in many other studies) has always been difficult to explain in terms of a maturity principle. How is becoming closed to new ideas and differences a sign of increased productive maturity? In summary, to the extent that clear predictions based on the maturity principle can be made a priori, the results of the present investigation do not seem to be fully consistent with it, especially regarding old age and the specific results obtained for Extraversion and Openness. In terms of a priori explanatory power, the maturity principle has thus limited usefulness to understanding changes in FFA factors with age. At best, our research – consistent with other research – suggests that support for the productive-maturity effect is limited to the late-adolescent to early-adult period.

La Dolce Vita effect. Clearly there is no support for the plaster hypothesis or the productive maturity effect in old age. Indeed, the term “maturity” does not even seem to make sense for the elderly. Based on self-concept research, we suggested that – despite obvious declines in particular physical attributes – the elderly tend to become more content with themselves in old age as reflected in higher levels of self-esteem. We labeled this the La Dolce Vita effect. The results of the present investigation seem to be consistent with the emergence of such self-contentment in old age as people become happier (more agreeable, less neurotic), more self-content and self-centered (less extroverted and open), more laid back and satisfied with what they have (less conscientious, open, outgoing and extroverted), and less preoccupied with productivity. This seems to suggest that, with age, after having devoted their lives to work, career and family, people tend to embrace more positive attitudes towards life and maybe to embrace more positively what life still has in store for them, personally. Interestingly, our results based on FFA factors apparently converge with other studies we reviewed that considered changes in personality that emerge after the age of 55 (e.g. Donellan, & Lucas, 2008; Roberts et al., 2006a, b; Terracciano et al., 2005).

As FFA factors have been purported to represent the core of human identity (e.g., Boyle, 2008; Caspi et al., 2005; Digman, 1990; Marsh, Trautwein, et al., 2006), similar effects should be observed in other domains and indeed, this seems to be the case, thus reinforcing the La Dolce Vita proposition. First, as we previously noted, Marsh, Martin and Jackson (2010) observed that as people get older, they become aware of their declining physical attributes, but also becoming more accepting of this decline, potentially due to a reduced incorporation of external frames of references in their internal standards (for a similar interpretation in the psychodynamic area, see Hillman, 1999). Moreover, indeed, these authors observed that for global self-esteem and global physical self-concept, older people tend to become more satisfied with themselves overall even though they decline in relation to particular physical attributes. This appears to be linked to the incorporation of more efficient self preservation and adaptive strategies in older age (e.g. Alaphilippe, 2008; Brandtstädter & Greve, 1994; Carstensen & Freund, 1994). For instance, Brandtstädter and Greve (1994) describe the self-concepts of older adults as characterized by an increased level of resourcefulness and flexibility due to the action of three interrelated mechanisms aiming at: (a) preventing or compensating losses in domains that are central to the identify; (b) readjusting personal goals or aspirations to avoid negative self-evaluations; (c) immunizing self identity against contradictory evidence through, for instance, selective perception (i.e., reduced openness).

Support for the La Dolce Vita effect also comes from previous research showing that older people report fewer negative interpersonal interactions than younger people (i.e., are more agreeable) and that when they do, they also report less negative affect (e.g. Almeida, 2005; Birditt & Fingerman, 2005; Lefkowitz &

Fingerman, 2003). In a study designed to investigate the reasons behind this observation, Charles and Carstensen (2008) reported that, when facing negative social interactions, older adults made less negative comments about the speaker, but also made fewer appraisals about them generally and expressed less desire to learn about their motives. They thus seem to simply disengage from these interactions (i.e. they are less open/extraverted). All of these results suggest that the La Dolce Vita interpretation provided here may represent a central mechanism to a positive human aging process in which people strive to devote their energy to the enjoyment of what life still has to offer whilst relying on more efficient self-regulatory mechanisms which allows them to devote less energy to unpleasant experiences. This could also reflect the observed decrease in openness as familiarity tends to require less efforts of adaptation.

Finally the La Dolce Vita effect is also generally consistent with evidence from psychiatric epidemiology showing that the point prevalence and/or severity of most forms of mood, anxiety, behavioural, substance abuse-related and personality disorders tend to decrease past 50 or 60 years old, with very few new onsets occurring after that time (e.g. Degenhardt et al., 2008; Grant et al., 2004; Huang et al., 2009; Jackson, & Burgess, 2000; Kessler, et al., 2005, 2007; Lenzenweger, et al., 2007; The ESEMeD ⁄MHEDEA-2000 Investigators, 2004a, 2004b). Indeed, even depression, which was long thought to increase with age, was in fact found to decrease once physical health and illnesses were controlled (Kessler et al., Birnbaum, Bromet, Hwang, Sampson, & Shahly, 2010; Kessler, Birnbaum, Shaly, et al., 2010). Although the observed decrease in anxiety may in appearance contradict the current results showing a similar decline in Openness, this juxtaposition suggests a tendency to avoid anxiety-generating situations through the aforementioned self-preservation mechanisms. Thus, there is an increase in Agreeableness and a decrease in Neuroticism (increase in emotional stability) in part because there is a decrease in Openness.

In summary, our results show that there are systematic developmental changes in personality over the entire adolescent and adult life-span, leading us to reject the plaster hypotheses. Although there is some support for a “productive-maturation” effect in the the adolescent to early-adult period, this support does not generalize to old age. Particularly as individuals grow into old age, they seem to reach a point of contentment that we have characterised as the La Dolce Vita effect. Although our introduction of the term La Dolce Vita effect in the present investigation is speculative, the effect appears to bring together patterns of changes in old age from a variety of different psychological disciplines. Further research is warranted to explore this effect and the reasons behind it.

Gender differences. Based on reviews of the FFA literature, women tend to score higher than men on Neuroticism and Agreeableness and perhaps also on Conscientiousness and Extraversion. However, there are no clear trends in gender differences for Openness. Our results are reasonably consistent with these expectations. The major differences are that we found almost no gender differences in Conscientiousness, whilst men had substantially higher scores on Openness than women. Perhaps the most striking finding of our study was how remarkably consistent these gender differences were across such a wide age range (15-99). Although there were systematic age differences, these changes as a function of age were nearly the same for men and women. ESEM models that constrained age-by-gender interactions to be zero fit the data nearly as well (in some cases better according to fit indexes that control for parsimony) than models where these interaction effects were freely estimated.

Summary of Methodological Implications.

Multiple-Group-MIMIC Hybrid. Latent variable analysts have typically used two main approaches to testing mean differences across groups: Multiple group comparisons and the MIMIC models. Both these approaches have critical, counter-balancing strengths and weaknesses. The MIMIC model is much more parsimonious and thus more attractive to applied studies– particularly those based on modest sample sizes. Importantly, the MIMIC approach is equally appropriate to truly categorical variables (e.g., gender), continuous variables (e.g., age), or a mixture of the two (as in the present investigation). However, critical assumptions of measurement invariance are implicit in the MIMIC model and cannot be tested. The particular strength of the multiple group models is that it allows tests of the full range of invariance tests like those considered here. Many multiple group comparisons are based on only two groups (or a small number of groups representing different levels of a single variable). However, we demonstrated here that this could easily be expanded to include all levels of two or more variables and their interactions (i.e., the six age-gender groups representing all combinations of the 3 age x 2 gender groups considered here). Major limitations of the multiple-group approach are the very large number of estimated parameters (which typically require large Ns) and the assumption that all variables of interest can be represented by a small number of categories. Although some variables (e.g., gender) are naturally categorical, many are not. In psychological research, it is well known that there are serious limitations in using a small number of categories to represent a reasonably continuous variable like age (MacCallum, et al., 2002). Hence, both the multiple-group and MIMIC models are likely to be “wrong” for different reasons. In the present investigation, we explored a hybrid approach that incorporated advantages of both the MIMIC and multiple group approaches. Again we note that this hybrid approach could be applied with either ESEM or CFA models, but CFA models would be inappropriate (as would corresponding analyses based on manifest variables) if the CFA models did not adequately fit the data or the fit of ESEM models was substantially better. This application of the hybrid multiple-group MIMIC approach makes three main contributions. First, we independently applied both the MIMIC and multiple group approaches to the same data. Particularly important were the tests of invariance (or partial invariance) in the multiple group approach that was implicit in the MIMIC approach. Results from these two contrasting approaches provided very similar results. Second, here we expanded the use of this hybrid approach by actually incorporating both approaches into a single model so that age and gender effects are based on both approaches that resulted in a graph that incorporated both multiple group and MIMIC effects of gender and age. Third, this application demonstrates the flexibility of the ESEM approach. We also note that other combinations of MIMIC and multiple group models are possible. For example, it would have been possible to treat only gender as a multiple group variable and age as a MIMIC variable. Although less complete than the models investigated in the present investigation, these alternative models may prove quite useful with smaller samples sizes.

ESEM vs. CFA. Why have FFA researchers not taken more advantage of the tremendous advances in statistical methodology that appear to be highly relevant to important substantive concerns like those considered here? Many of these advances are based substantially on CFA and related statistical techniques. Marsh, et al. (2009, Marsh, Lüdtke et al., 2010) argued that the traditional ICM-CFA model is not appropriate for many well-established psychological measures, including most FFA measures. Indeed, this view is commonly expressed by FFA researchers (e.g., McCrae et al., 1996). However, personality researchers proclaiming the inappropriateness of CFA were 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. In at least some situations, as demonstrated here, this apparent impasse can be largely overcome through application of ESEM. 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 traditional ICM-CFA model, but we do not argue that the ESEM approach should replace the CFA approach. Indeed, when the more parsimonious ICM-CFA model fits the data as well as the ESEM model and results in similar parameter estimates, the ICM-CFA should be used. However, when the ICM-

CFA model is unable to fit the data whilst 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.

In summary: (a) responses to FFA instruments (but, more generally, most psychological measures) typically do not meet the assumptions of the ICM-CFA model and will result in biased estimates if used despite these problems; (b) if the ESEM model fits the data better than the ICM-CFA model, then the assumptions of the ICM-CFA model are unlikely to be valid; (c) in many instances, the less restrictive assumptions of the ESEM model provide more valid estimates.

FFA research has largely ignored fundamental issues related to complex structures of measurement error. Although FFA researchers routinely report coefficient alpha estimates of reliability, the “state of the art” has moved well beyond these historically acceptable measures. Simply reporting coefficient alpha estimates of reliability provides an index of one aspect of measurement error, but largely ignores other aspects of unreliability and does 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, Seaton et al., 2010). The ability to define and control for complex structures of measurement error has been one of the important advances available to applied researchers through the application of CFA, but these advances are largely absent in traditional approaches to EFA. An important advantage of ESEM is to provide many of the advantages of CFA without the constraints imposed by the traditional ICM-CFA factor structure. Although ESEM does not allow the full flexibility of CFA/SEM models as currently operationalized in Mplus (e.g., constraints on group specific correlations among factors, tests of higher-order factor models, fully latent curve models, factor mixture models, etc.), we also proposed (see appendix) an extension of ESEM models (ESEM-Within-CFA) that can be used to circumvent most of these current limitations.

Methodological limitations and directions for further research.

Reliance on cross-sectional data. An important limitation of the present investigation is reliance on cross-sectional data – particularly in relation to chronological age – that require additional caveats in the interpretation of the results. For example, the apparent differences as a function of age – particularly in old age – could reflect relations of FFA factors with longevity or mortality (see related discussion by Caspi et al., 2005). Similarly, it is important to acknowledge that observed differences may also be a function of birth cohort effects (see related discussion by Roberts, et al., 2006a, and Twenge, 2000, 2001). Reliance on cross-sectional data thus limits the issues that we were able to address. Thus, we were not able to evaluate how consistent changes in FFA factors were for different individuals, as this would have required longitudinal data (also see discussion by Block, 2010 of person-centered approaches). For example, Costa et al. (1999) proposed an extended version of the plaster hypothesis, suggesting that in addition to mean-level stability, FFA traits were also characterized by rank-order stability over time (i.e. by stable inter-individual differences). Although our results are clearly inconsistent with the plaster hypothesis in relation to mean level differences, we were not able to examine the stability of individual differences with our cross-sectional data.

Although there are many advantages for longitudinal data, there are also some limitations. To the extent that the data are based on a single age cohort, then there are issues about the generalizability of the results to other age cohorts. Problems associated with mortality and longevity also affect longitudinal data, although longitudinal data provide a stronger basis for evaluating the consequences of these issues. Particularly for large, nationally representative samples, longitudinal data is much more expensive and time-consuming to collect and more likely to be plagued non-random missing data. Furthermore, it would not be realistically possible to collect a longitudinal dataset that covered the range of ages (15 to 100) covered here. The best possible compromise would be a multicohort multiwave design that combines advantages of both longitudinal and cross-sectional data. However, even here there would still be the problem of cohort and mortality variations with the older cohorts. Although there is no solution to this problem, at least our sample is a nationally representative sample of people who are currently alive that covers one of the most extensive age-ranges ever considered in FFA studies. . Ultimately the “best” description of how FFA factors change with age must be able to incorporate findings from both cross-sectional and longitudinal studies. We also note that it would be possible to evaluate true longitudinal data with ESEM (see Marsh, Lüdtke et al., 2010), to test the invariance of responses using essentially the same set of invariance models considered here, and to compare ESEM results with those based on traditional ICM-CFA approaches.

Limitations in the Applications of ESEM. ESEM is a relatively new statistical tool and the development of best practice will have to evolve with experience and application. Limitations and directions for further research are discussed in more detail elsewhere (Asparouhov & Muthén, 2009; Marsh et al., 2009; Marsh, Lüdtke et al., 2010). Particularly relevant to the present investigation are issues related to goodness of fit assessment, the appropriateness of partial invariance models based on ex post facto modifications, and analyses based on responses to individual items. Some of these issues are overcome by the application of our taxonomy of models focusing on the relative fit of competing models. However, we recommend that researchers use an eclectic approach based on a subjective integration of a variety of indexes, 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. The use of ex post-facto modification indexes to construct models of partial invariance in ESEM is worrisome but applies to CFA studies as well. Without softening invariance assumptions to include partial invariance (e.g., invariance of intercepts in gender and age groups), the applied researcher is not entitled to pursue substantive questions of interest. Whilst it might be possible to develop better instruments that are more fully invariant, we suspect that this will continue to be an ongoing issue in applied research.

We also note that partial invariance models are clearly more defensible than analyses based on manifest scores that implicitly assume complete invariance. In the present investigation, we started at the item level. Some researchers have attempted to circumvent concerns related to CFA and partial invariance through the use of item aggregates: facet scores (e.g., Ashton et al., 2009; Gignac, 2009; McCrae et al., 1996; Saucier, 1998; Small et al., 2003), parcels (e.g., Allemand et al., 2007, 2008; Lüdtke et al., 2009; Marsh et al., 2006), or scale scores (e.g., Mroczek & Spiro, 2003). Although potentially appropriate and useful for some specific purposes, the use of item aggregates – by definition – does not allow researchers to test appropriately differential item functioning and measurement invariance at the level of the individual items. Furthermore, unless very strict assumptions are met, analyses based on aggregates of items are likely to camouflage misfit at the item level and result in biased parameter estimates and relations among factors (e.g., Bandalos, 20008; Marsh, et al., 2011). Indeed, Marsh, et al., argue that unless the ICM-CFA model fits the data as well as ESEM models, there are potentially serious violations of assumptions of unidimensionality upon which parceling strategies are based. Hence, we recommend that applied researchers who chose to do CFA (or ESEM) analyses at the item-aggregate level should also evaluate the appropriateness of their models and interpretations at the individual item level.

How well can the FFA factors be explained in terms of only 15 items? Is this FFA instrument simply too short? Short forms are very controversial (Marsh, Ellis, Parada, Richards & Heubeck, 2005) and even the widely used 60-item NEO-FFI is a compromise “short” version of longer (180- and 270-item) instruments.

This is an important issue as increasingly, FFA researchers recognize that results as basic as gender and age differences depend in part on the items (or subfacets) used to measure the FFA factors (McCrae & Costa, 2001; Terracciano et al., 2005). Ultimately this is an issue of differential item functioning that is most appropriately addressed through tests of measurement invariance like those pursued here. However, these tests of invariance relate to the generality of findings across the items that were considered, not how the items used in a particular study map onto the population of items that could have been used (or samples of item used on other instruments). One consequence of measuring FFA factors with so few items is the inevitably low levels of reliability (see earlier discussion). We note, however, that this is not an inherent problem so long as latent variables models are used to correct for unreliability, as in the present investigation. Clearly, the use of such an abbreviated FFA instrument is an expedient compromise that made it possible for FFA measures to be included in the British Household Panel Survey.

In summary, ESEM is not a panacea and may not be appropriate in some situations. However, it provides developmental and personality researchers with considerable flexibility to address substantively important 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 evolve with experience. Nevertheless, results of the present investigation (also see Asparouhov & Muthén, 2009; Marsh et al., 2009; Marsh, Lüdtke et al., 2010) provide strong support for the application of ESEM in psychological research more generally.

Footnotes

1 – The four non-invariant items, and the intercepts for males (M) and females (F), were: Openness item 2 (M=4.16, F=4.61), Agreeableness item 1 (M=5.64, F=6.00), Conscientiousness item 1 (M=5.27, F=5.08), and Conscientiousness item 2 (M=5.08, F=5.32). See Table 2 for wording of items.

References

Alaphilippe, D. (2008). Evolution de l'estime de soi chez l'adulte âgé [Self-esteem in the elderly]. *Psychologie et Neuropsychiatrie du Viellissement*, *6*, 167-176.

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.

Almeida, D.M. (2005). Resilience and vulnerability to daily stressors assessed via diary methods. *Current directions in Psychological Sciences, 14,* 64-68.

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

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

Bagozzi, R. P. (1993). Assessing construct validity in personality research: Applications to measures of self-esteem. *Journal of Research in Personality, 27,* 49-87.

Bandalos, D. L. (2008). Is parceling really necessary? A comparison of results from item parceling and categorical variable methodology. Structural Equation Modeling, 15(2), 211-240.

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.

Birditt, K.S., & Fingerman, K.L. (2005). Do we get better at picking our battles? Age group differences in descriptions of behavioural reactions to interpersonal tensions. *Journal of Gerontology: Psychological Sciences*, *60B*, 121-128.

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

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.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*, 425-440.

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.

Brandtstädter, J., & Greve, W. (1994). The aging self: stabilizing and protective processes. *Developmental Review*, *14*, 52-80.

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.

Cartesen, L.L., & Freund, A.M. (1994). The resilience of the aging self. *Developmental Review*, *14*¸81-92.

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.

Charles, S.T., & Carstensen, L.L. (2008). Unpleasant situations elicit different emotional responses in younger and older adults. *Psychology and Aging*, *23,* 495-504.

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. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9,* 233-255.

Church, A. T., & Burke, P. J. (1994). Exploratory and Confirmatory Tests of the Big 5 and Tellegens 3-Dimensional and 4-Dimensional Models. *Journal of Personality and Social Psychology, 66,* 93-114.

Corwyn, R. F. (2000). The factor structure of global self-esteem among adolescents and adults. *Journal of Research in Personality*, *34*, 357–379.

Costa, P. T., & 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 Jr., P. T., & McCrae, R. R. (1994). Set like plaster? Evidence for the stability of adult personality. In T. F. Heatherton & J. L. Weinberger (Eds.), Can personality change? (pp. 21-40). Washington, DC, US: American Psychological Association.

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 Jr., P. T., & McCrae, R. R. (1997). Stability and change in personality assessment: The Revised NEO Personality Inventory in the Year 2000. *Journal of Personality Assessment, 68*, 86-94.

Costa Jr., P. T., McCrae, R. R., & Siegler, I. C. (1999). Continuity and change over the adult life cycle: Personality and personality disorders. In C. R. Cloninger (Ed.), *Personality and psychopathology*. (pp. 129-154). Washington, DC, US: American Psychiatric Association.

Costa, P., 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.

de Raad, B., & Hofstee, W. K. (1993). A circumplex approach to the Five Factor model: A facet structure of trait adjectives supplemented by trait verbs. Personality and Individual Differences, 15(5), 493-505

Degenhardt, L., Chiu, W.-T., Sampson, N., Kessler, R.C., Anthony, J.C., Angermeyer, M., Bruffaerts, R., de Girolamo, G., Gureje, O., Huang, Y., Karam, A., Kostyuchenko, S., Lepine, J.P., Medina Mora, M.E., Neumark, Y., Ormel, J.H., Pinto-Meza, A., Posada-Villa, J., Stein, D.J., Takeshima, T., & Wells, J.E. (2008). Toward a global view of alcohol, tobacco, cannabis, and cocaine use: Findings from the WHO World Mental Health Surveys. *PLoS Medicine, 5*, 1053-1067.

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology*, *41*, 417-440.

Dolan, C.V., Oort, F.J., Stoel, R.D., & Wicherts, 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.

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

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

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

Grant, B.F., Hasin, D.S., Stinson, F.S., Dawson, D.A., Chou, S.P., Ruan, W.J., & Pickering, R.P. (2004). DSM-IV Personality Disorders in the National Comorbidity Survey Replication. *Journal of Clinical Psychiatry*, *65*, 948-958.

Guo, S. (1995) Sex Differences in Personality: A Meta-Analysis Based on "Big Five" Factors. Paper presented at the Annual Meeting of the AERA (San Francisco, CA. ERIC Document Number ED383759).

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: Johns Hopkins University Press.

Helson, R., Kwan, V. S.Y., John, O. & Jones, C. (2002). The growing evidence for personality change in adulthood: findings from research with personality inventories. *Journal of Research in Personality*, *36*, 287–306

Helson, R., & Malone, G. (1987). Personality-change in women from college to midlife. *Journal of Personality & Social Psychology, 53*, 176-186

Hessen, D. J., Dolan, C. V., & Wicherts, J. M. (2006). The Multigroup Common Factor Model With Minimal Uniqueness Constraints and the Power to Detect Uniform Bias. Applied Psychological Measurement, 30(3), 233-246.

Hillman, J. (1999). *The Force of Character: And the Lasting Life*. New York, NY: Random house.

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.

Huang, Y., Kotov, R., de Girolamo, G., Preti, A., Angermeyer, M., Benjet, C., Demyttenaere, K., de Graaf, R., Gureje, O., Nasser Karam, A., Lee, S., Le Pine, J.P., Matschinger, H., Posada-Villa, J., Suliman, S., Vilagut, G., & Kessler, R.C. (2009). DSM–IV personality disorders in the WHO World Mental Health Surveys. *The British Journal of Psychiatry*, *195*, 46-53.

Jackson, H.J., & Burgess, P.M. (2000). Personality disorders in the community: A report from the Australian National Survey of Mental Health and Wellbeing. *Social Psychiatry and Psychiatric Epidemiology, 35,* 531-538.

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 multi-method, multi-sample study of age differences in the facets of conscientiousness. *Journal of Personality and Social Psychology, 96*, 446-459.

James, W. (1890/1963). *The principles of psychology* (Vol. 2). New York: Holt.

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: Guilford.

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: Academic Press.

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

Kessler, R.C., Angermeyer, M., Anthony, J.C., de Graaf, R., Demyttenaere, K., Gasquet, I., de Girolamo, G., Gluzman, S., Gureje, O., Haro, J.M., Kawakami, N., Karami, A., Levinson, D., Medina Mora, M.E., Oakley Browne, M.A., Posada-Villa, J., Stein, D.J., Adley Tsang, C.H., Aguilar-Gaxiola, S., Alonso, J., Lee, S., Heeringa, S., Pennell, B.-E., Berglund, P., Gruber, M.J., Petukhova, M., Chatterji, S., & Üstün, T.B. (2007). Lifetime prevalence and age-of-onset distributions of mental disorders in the World Health Organization's World Mental Health Survey Initiative. *World Psychiatry, 6*, 168-176.

Kessler, R.C., Berglund, P., Demler, O., Jin, R., & Walters, E.E. (2005). Lifetime Prevalence and age-of-onset Distributions of DSM-IV disorders in the National Comorbidity Survey replication. *Archives of General Psychiatry*, *62*, 593-602.

Kessler, R.C., Birnbaum, H., Bromet, E., Hwang, I., Sampson, N., & Shahly, V. (2010). Age differences in major depression: Results from the National Comorbidity Survey Replication (NCS-R). *Psychological Medicine*, *40*, 225-237.

Kessler, R.C., Birnbaum, H., Shahly, V., Bromet, E., Hwang, I., McLaughlin, K.A., Sampson, N., Andrade, L.H., de Girolamo, G., Demyttenaere, K., Haro, J.M., Karam, A.N., Kostyuchenko, S., Kovess, V., Lara, C., Levinson, D., Matschinger, H., Nakane, Y., Oakley Browne, M., Ormel, J., Posada-Villa, J., Sagar, R.,J.., & Stein, D. (2010). Age differences in the prevalence and co-morbidity of DSM-IV major depressive episodes: Results from the WHO World Mental Health Survey Initiative. *Depression and Anxiety*, *27*, 351-364.

Klimstra, T. A., Hale III, W. W., 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.

Lefkowitz, E.S., & Fingerman, K.L. (2003). Positive and negative emotional feelings and behaviors in mother-daughter ties in late life. *Journal of Family Psychology*, *17*, 607-617.

Lenzenweger, M.F., Lane, M.C., Loranger, A.W., & Kessler, R.C. (2007). DSM-IV Personality Disorders in the National Comorbidity Survey Replication. *Biological Psychiatry*, *62*, 553-564.

Lüdtke, Trautwein & Husemann (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

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7,* 19-40.

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

Marsh, H. W. (1986). The bias of negatively worded items in rating scales for young children: A cognitive-developmental phenomena. *Developmental Psychology, 22,* 37-49.

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

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

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 on sport psychology* (3rd ed., pp. 774 - 798). New York: Wiley.

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., Balla, J. R., & Hau, K. T. (1996). An evaluation of incremental fit indexes: 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., Ellis, L., Parada, L., Richards, G. & Heubeck, B. G. (2005). A short version of the Self Description Questionnaire II: Operationalizing criteria for short-form evaluation with new applications of confirmatory factor analyses. *Psychological Assessment, 17,* 81-102.

Marsh, H. W., & Grayson, D. (1994). Longitudinal stability of latent means and individual differences: A unified approach. *Structural Equation Modeling, 1,* 317-359.

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*. 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., Lüdtke, O., Muthén, B., Asparouhov, T., Morin, A. J. S., Trautwein, U., & Nagengast, B. (2010). A new look at the big-five factor structure through exploratory structural equation modeling. *Psychological* Assessment, *22*, 471-491.

Marsh, H. W, Lüdtke,O., Robitzsch, A., Nagengast, B., Morin, A. J. S. & Trautwein, U. (2011, in review). Two wrongs do not make a right: Camouflaging misfit with item-parcels in cfa models.

Marsh, H. W., Martin, A. J. & Jackson, S. (2010). Introducing A short version of the Physical Self Description Questionnaire: New strategies, short-form evaluative criteria, and applications of factor analyses. *Journal of Sport & Exercise Psychology,32*, 438-482.

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., Scalas, L. F. & Nagengast, B. (2010). Longitudinal tests of competing factor structures for the rosenberg self-esteem scale: Traits, ephemeral artefacts, and stable response styles. Psychological Assessment,22, 366-381.

Marsh; H. W., Seaton, M., Kuyper, H., Dumas; F., Huguet, P., Regner, I., Buunk, A. P., Monteil, J. M, Blanton, H., Gibbons, F. X. (2010). Phantom behavioral assimilation effects: Systematic biases in social comparison choice studies. *Journal of Personality, 78*, 671-710.

Marsh, H. W., Tracey, D. K., Craven, R. G. (2006). Multidimensional self-concept structure for preadolescents with mild intellectual disabilities: A hybrid multigroup-mimic approach to factorial invariance and latent mean differences. *Educational and Psychological Measurement, 66,* 795-818.

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., 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.

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. (2001). Gender differences in personality traits across cultures: Robust and surprising findings. *Journal Of Personality And Social Psychology*, *81*, 322-331.

McCrae, R. R., Costa Jr., P. T., Ostendorf, F., Angleitner, A., Hrebícková, M., Avia, M. D., Sanz, J., Sánchez-Bernardos, M. L., Kusdil, M. E., Woodfield, R., Saunders, P. R., & Smith, P. B. (2000). Nature over nurture: Temperament, personality, and life span development. *Journal of Personality and Social Psychology, 78*, 173-186.

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*, *58B*, 153-165.

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

Muthén, L. K., & Muthén, B. (2009). *Mplus Short courses. Topic 1:* Exploratory Factor Analysis, Confirmatory Factor Analysis, And Structural Equation Modeling For Continuous Outcomes. Los Angeles CA: Muthén & Muthén. [http://www.statmodel.com/download/Topic%201-v11.pdf]

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.

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. Journal of Research in Personality, *41*, 203-212.

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

Roberts, B. W., Walton, K. E., & Viechtbauer, W. (2006b). Personality traits change in adulthood: Reply to Costa and McCrae (2006). *Psychological Bulletin*, *132*, 29-32.

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.

Schmitt, T.A., & Sass, D.A. (2011). Rotation criteria and hypothesis testing for exploratory factor analysis: Implications for factor pattern loadigns and interfacto correlations. *Educational and Psychological Measurement*, *71*, 95-113.

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*, *58B*, 166–176.

Srivastava, S., John, O. P., Gosling, S. D., & Potter, J. (2003). Development of personality in early and middle adulthood: Set like plaster or persistent change? *Journal of Personality and Social Psychology*, *84*, 1041-1053.

Taylor, M. F., Brice, J., Buck, N. & Prentice-Lane, E. (2009) *British Household Panel Survey User Manual Volume A: Introduction, Technical Report and Appendices.* Colchester, UK: University of Essex.

Terracciano, A., McCrae, R. R., Brant, L. J., & Costa Jr., P. T. (2005). Hierarchical linear modeling analyses of the NEO-PI-R Scales in the Baltimore Longitudinal Study of Aging. *Psychology and Aging, 20,* 493-506.

The ESEMeD⁄MHEDEA 2000 Investigators (2004a). Prevalence of mental disorders in Europe: results from the European Study of the Epidemiology of Mental Disorders (ESEMeD) project. *Acta Psychiatrica*

Scandinavica, *109* (Suppl. 420), 21-27.

The ESEMeD⁄MHEDEA 2000 investigators (2004b). 12-Month comorbidity patterns and associated factors in Europe: results from the European Study of the Epidemiology of Mental Disorders (ESEMeD) project. *Acta Psychiatrica Scandinavica*. *109* (Suppl. 420), 28-37.

Twenge, J. M. (2000). The age of anxiety? The birth cohort change in anxiety and neuroticism, 1952–1993. Journal of Personality and Social Psychology, 79, 1007-1021.

Twenge, J. M. (2001). Birth cohort changes in extraversion: A cross-temporal meta-analysis, 1966–1993. Personality and Individual Differences, 30(5), 735-748.

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.

Table 1

Taxonomy of Invariance Tests Designed to Evaluate 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, 2, 3, 4]
7	FL, Uniq, Inter [1, 2, 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, 2, 3, 5, 7, 10] (Manifest mean invariance)
12	FL, FVCV, Inter, LFMn [1, 2, 4, 5, 6, 8, 10]
13	FL, Uniq, FVCV, Inter, LFMn [1-12] (complete factorial invariance)

Note. FL= Factor Loadings; FVCV=Factor variance-covariances; Inter = item intercepts; Uniq = item uniquenesses; LFMn = Latent Factor Means. Models with latent factor means freely estimated constrain intercepts to be invariant across groups, whilst models where intercepts are free imply that mean differences are a function of intercept differences. Brackets values 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) whilst model 13 (complete invariance) is nested under all other models. Parts of this table were adapted from 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,* Table 1, p. 443).

Table 2

Summary of Goodness of Fit Statistics for Total Group Models

CHI	df	CFI	TLI	RMSEA	Description
Total Group—Big Five Only					
Total Group CFA					
6629	80	.761	.687	.076	TGCFA1A: no CUs
5455	74	.804	.722	.072	TGCFA1B: CUs
Total Group ESEM					
1200	40	.958	.889	.045	TGESEM1A: no CUs
497	34	.983	.948	.031	TGESEM1B: CUs
Total Group MIMIC—Age (L=linear &Q= quadratic)					
1069	54	.966	.916	.037	MIMICAge1, Age (L & Q) full intercepts invariance
779	51	.976	.936	.032	MIMICAge2, Age (L & Q), partial intercepts invariance
1149	56	.964	.912	.037	MIMICAge3, Age (L, Q = 0), partial intercepts invariance
Total Group MIMIC—Age (L & Q), Sex, and Age (L & Q) by Sex interactions					
1209	81	.966	.924	.032	MIMICAge*sex1
927	77	.974	.940	.028	MIMICAge*sex2, partial intercepts invariance
973	87	.973	.945	.027	MIMICAge*sex3, partial intercepts invariance, interaction fixed to 0.

Note. CHI= chi-square; df=degrees of freedom ratio; CFI= Comparative fit index; TLI=Tucker-Lewis Index; RMSEA= Root Mean Square Error of Approximation. CUs = a priori correlated uniquenesses based on the negatively worded items. All analyses were weighted by the appropriate weighting factor and based on a complex design option to account for nesting within families

Table 3

Factor Solutions: Five-Factor CFA and ESEM Solutions Based on Responses to 15 items

	CFA (TGCFA1B in Table 2)						ESEM (TGESEM1B in Table 2)						
Factor Loadings													
	Agre	**Conc**	**Extr**	**Neur**	**Open**	**R-Sq**	**Agre**	**Conc**	**Extr**	**Neur**	**Open**	**R-Sq**	**Item Wording**
F1 Agreeableness (A)													I see myself as someone who:
A1R	.27	.00	.00	.00	.00	.07	.40	-.04	-.11	-.15	-.17	.19	Is sometimes rude to others (reverse-scored).
A2	.49	.00	.00	.00	.00	.24	.47	.03	.10	.05	.12	.30	Has a forgiving nature.
A3	.92	.00	.00	.00	.00	.84	.69	.23	.09	.06	.00	.70	Is considerate and kind to almost everyone.
F2 Conscientiousness (C)													
C1	.00	.54	.00	.00	.00	.29	-.04	.51	.19	.05	.13	.38	Does a thorough job.
C2R	.00	.31	.00	.00	.00	.09	.03	.34	-.05	-.19	-.21	.17	Tends to be lazy (reverse-scored).
C3	.00	.88	.00	.00	.00	.78	.21	.72	.02	-.02	.05	.72	Does things efficiently.
F3 Extraversion (E)													
E1	.00	.00	.63	.00	.00	.40	.03	.09	.74	.09	.03	.60	Is talkative.
E2	.00	.00	.82	.00	.00	.67	.20	.03	.56	-.11	.14	.49	Is outgoing, sociable.
E3R	.00	.00	.24	.00	.00	.06	-.15	-.23	.39	-.21	-.10	.26	Is reserved (reverse-scored).
F4 Neuroticism (N)													
N1	.00	.00	.00	.72	.00	.53	.01	.06	.06	.75	-.00	.56	Worries a lot.
N2	.00	.00	.00	.69	.00	.48	.11	-.05	-.08	.68	.05	.48	Gets nervous easily.
N3R	.00	.00	.00	.55	.00	.30	-.19	-.14	-.04	.51	-.19	.42	Is relaxed, handles stress well (reverse-scored).
F5 Openness (O)													
O1	.00	.00	.00	.00	.70	.49	-.07	.14	.10	-.05	.65	.53	Is original, comes up with new ideas.
O2	.00	.00	.00	.00	.55	.31	.09	-.03	.04	.08	.56	.35	Values artistic, aesthetic experiences.
O3	.00	.00	.00	.00	.68	.46	.10	.10	.10	-.06	.55	.43	Has an active imagination, is original, comes up with new ideas.
Factor Correlations													
A	1.0						1.0						
C	.68	1.0					.41	1.0					
E	.42	.41	1.0				.15.	.19	1.0				
N	.03	-.09	-.17	1.0			-.01	-.05	-.07	1.0			
O	.35	.47	.56	-.08	1.0		.16	.25	.31	.01	1.0		

Note. The CFA and ESEM models each specified five factors (see Table 2 for goodness of fit statistics). All parameter estimates are completely standardized. N = 14,932 sets of ratings for the 15 big-five items. Both models also included a set of a priori correlated uniquenesses, relating negatively worded items

Table 4

Summary of Goodness of Fit Statistics for Gender Invariance Models

CHI	df	CFI	TLI	RMSEA	Description
Model MG1: (configural invariance)					
1234	80	.958	.889	.045	MG1A: no invariance (configural invariance)
580	68	.981	.942	.033	MG1B: MG1A with CUs (not invariant over sex)
616	74	.980	.944	.032	MG1C: MG1B with CUs IN (invariant over sex)
Model MG2(FL, weak factorial/measurement invariance)					
1346	130	.956	.928	.037	MG2A
750	118	.977	.959	.028	MG2B: MG2A with CUs
765	124	.977	.960	.027	MG2C: MG2B with CUs IN
Model MG3 (FL & Uniq)					
1456	145	.952	.931	.036	MG3A
868	133	.973	.958	.028	MG3B: MG3A with CUs
882	139	.973	.959	.028	MG3C: MG3B with CUs IN
Model MG4 (FL & FVCV)					
886	139	.973	.959	.028	MG4C: MG4 with CUs IN
Model MG5 (FL & Inter; Strong factorial/measurement invariance)					
1076	134	.966	.946	.032	MG5C: MG5 with CUs IN
802	130	.975	.960	.027	MG5Cp: MG5C, CUs IN, partial intercept invariance
Model MG6 (FL, FVCV, Uniq)					
1014	154	.969	.957	.028	MG6C: MG6 with CUs IN
Model MG7 (FL, Uniq, Inter; Strict factorial/measurement invariance)					
919	145	.972	.959	.028	MG7Cp: MG7 with CUs IN, partial intercept invariance
Model MG8 (FL, FVCV, Inter)					
922	145	.972	.959	.028	MG8Cp: MG8 with CUs IN, partial intercept invariance
Model MG9 (FL, Uniq, FVCV, Inter)					
1051	160	.967	.957	.028	MG9Cp: MG9 with CUs IN, partial intercept invariance
Model MG10 (FL, Inter, FMns; Latent mean invariance					
1978	135	.933	.895	.044	MG10Cp: MG10 with CUs IN, partial intercept invariance
Model MG11 (FL, Uniq, Inter, FMns; Manifest mean invariance)					
2083	150	.929	.901	.043	MG11Cp: MG11 with CUs IN, partial intercept invariance
Model MG12 (FL, FVCV, Inter, FMns)					
2086	150	.929	.901	.043	MG12Cp: MG12 with CUs IN, partial intercept invariance
Model MG13 (FL, Uniq, FVCV, Inter, FMns; complete factorial invariance)					
2200	165	.926	.905	.042	MG13Cp: MG13 MG9 with CUs IN, partial intercept invariance

Note. CHI= chi-square; df=degrees of freedom ratio; CFI= Comparative fit index; TLI=Tucker-Lewis Index; RMSEA= Root Mean Square Error of Approximation. CUs = a priori correlated uniquenesses based on the negatively worded items. All analyses were weighted by the appropriate weighting factor and based on a complex design option to account for nesting within families For multiple group invariance models, the "IN" means the sets of parameters constrained to be invariant across the multiple groups (P-IN means partial invariance): FL= Factor Loadings; FVCV=Factor variance-covariances; Inter = item intercepts; Uniq = item uniquenesses; FMn = Factor Means.

Table 5

Estimates of Age and Gender Effects in Big-Five Factors: MIMIC Models (also see See Table 2)

		MIMIC Models For Age Only						MIMIC Models for Age and Sex					
		MIMICAge1		MIMICAge2		MIMICAge3		MIMICAge*sex1		MIMICAge*sex2		MIMICAge*sex3	
		Est	SE	Est	SE	EST	AE	Est	SE	Est	SE	Est	SE
Agre	L-AGE	.08	.02	.10	.01	.09	.01	.09	.01	.08	.01	.08	.01
	Q-AGE	.03	.01	.04	.01	0	----	.04	.01	.03	.01	.03	.01
	SEX (S)							.18	.01	.13	.01	.13	.01
	S*L-AGE							.01	.01	.01	.01	0	----
	S*Q-AGE							-.02	.01	.02	.01	0	----
Conc	L-AGE	-.01	.02	-.06	.02	-.04	.01	-.06	.02	-.06	.02	-.06	.02
	Q-AGE	-.23	.02	-.23	.02	0	----	-.23	.02	-.23	.02	-.23	.02
	SEX							.01	.01	.06	.02	.06	.02
	S*L-AGE							-.05	.01	.05	.01	0	----
	S*Q-AGE							-.01	.01	.01	.01	0	----
Extra	L-AGE	-.28	.01	-.27	.01	-.27	.01	-.28	.01	-.28	.01	-.28	.01
	Q-AGE	.04	.01	.04	.01	0	----	.04	.01	.04	.01	.04	.01
	SEX							.15	.01	.17	.01	.17	.01
	S*L-AGE							-.05	.01	.04	.01	0	----
	S*Q-AGE							.01	.01	.01	.01	0	----
Neur	L-AGE	-.17	.01	-.22	.01	-.21	.01	-.23	.01	-.23	.01	-.23	.01
	Q-AGE	-.06	.01	-.06	.01	0	----	-.07	.01	-.07	.01	-.07	.01
	SEX							.31	.01	.32	.01	.32	.01
	S*L-AGE							-.00	.01	.00	.01	0	----
	S*Q-AGE							.01	.01	.01	.01	0	----
Open	L-AGE	-.30	.01	-.31	.01	-.31	.01	-.30	.01	-.30	.02	-.30	.02
	Q-AGE	-.03	.01	-.03	.01	0	----	-.02	.01	-.03	.01	-.03	.01
	SEX							-.17	.01	-.20	.01	-.20	.01
	S*L-AGE							.02	.01	.02	.01	0	----
	S*Q-AGE							.01	.01	.01	.01	0	----

Note. MIMIC = Mulitiple Indicator Multiple Cause. Est = unstandardized parameter estimate. SE = standard error. L-Age = linear component of age; Q-Age = Quadratic component of age; * = interaction effect. Based on a hierarchical design, the linear age component is the standardized (M=0, SD = 1) age, whilst the quadratic age component is the squared age component with the effect linear age partialed out (the quadratic component of age was not re-standardized so that it is in the same metric as the linear age component). Sex (-1 = male, +1 = female) was multiplied times the linear and age components to obtain the interaction terms. See Table 4 for a description of the six models and goodness-of-fit statistics.

Table 6
Multiple Group Invariance Tests: 6 (2 gender x 3 age) Multiple Age-Gender (MAG) Models

CHI	df	CFI	TLI	RMSEA	Description
Model MAG1: (configural invariance) [a]					
834	234	.979	.943	.034	MAG1, CUs invariant
Model MAG2(Factor Loadings=FL, weak factorial/measurement invariance)					
1566	484	.962	.950	.031	MAG2, CUs invariant
Model MAG3 (FL & Uniquenesses=UNQ) [b]					
3351	559	.901	.889	.046	MAG3, CUs invariant
1761	543	.957	.950	.031	MAG3p, CUs invariant and partial invariance of the uniquenesses
Model MAG4 (FL & factor variance-covariance = FVCV)					
2170	559	.943	.936	.035	MAG4, CUs invariant
Model MAG5 (FL & Intercepts=INT; Strong factorial/measurement invariance) [c]					
2542	535	.929	.917	.040	MAG5, CUs invariant
1722	514	.957	.948	.032	MAG5p, CUs invariant and partial invariance of the intercepts
Model MAG6 (FL, FVCV, UNQ)					
2397	618	.937	.936	.035	MAG6p , CUs invariant and partial invariance of the uniquenesses
Model MAG7 (FL, UNQ, INT; Strict factorial/measurement invariance)					
1904	572	.953	.948	.032	MAG7p, CUs invariant and partial invariance of the intercepts and uniquenesses
Model MAG8 (FL, FVCV, INT)					
2331	589	.938	.934	.036	MAG8p, CUs invariant and partial invariance of the intercepts
Model MAG9 (FL, UNQ, FVCV, INT)					
2538	647	.933	.935	.035	MAG9p, CUs invariant and partial invariance of the intercepts and uniquenesses
Model MAG10 (FL, INT, FMns; Latent mean invariance					
3961	539	.879	.859	.052	MAG10p, CUs invariant and partial invariance of the intercepts
Model MAG11 (FL, UNQ, INT, FMns; Manifest mean invariance)					
4123	597	.876	.869	.050	MAG11p, CUs invariant and partial invariance of the intercepts and uniquenesses
Model MAG12 (FL, FVCV, INT, FMns)					
4708	614	.855	.852	.053	MAG12p, CUs invariant and partial invariance of the intercepts
Model MAG13 (FL, UNQ, FVCV, INT, FMns; complete factorial invariance)					
4889	672	.851	.860	.052	MG13p, CUs invariant and partial invariance of the intercepts and uniquenesses

Note. CHI= chi-square; df=degrees of freedom ratio; CFI= Comparative fit index; TLI=Tucker-Lewis Index; RMSEA= Root Mean Square Error of Approximation. CUs = a priori correlated uniquenesses based on the negatively worded items. FL= Factor Loadings; FVCV=Factor variance-covariances; INT = item intercepts; UNQ = item uniquenesses; FMn = Factor Means.

In all analyses responses were weighted by the appropriate weighting factor and based on a complex design option to account for nesting within families. For multiple group invariance models, the “IN” means the sets of parameters constrained to be invariant across the multiple groups.

[a] In preliminary analyses, we found that correlated uniquenesses among negatively worded items were needed and that these were invariant over the six groups (see earlier discussion in relation to gender) and so all models presented in this table are based on this structure.[b] In models with invariances of uniqueness, additional models with partial invariance were tested. In the model MAG3p, 16 of 90 (15 items x 6 groups=90) uniqnesses were freed such that values for males and females in the oldest group were constrained to be equal to each other, but not to those from the other four groups. In addition, B5A1R was freed for females in the middle age group. This pattern of partial invariance of uniquenesses was used in all subsequent models with variance constraints on uniquenesses. [c] In models with invariances of intercepts, additional models with partial invariance were tested. In the model MAG5p 21 of 90 (15 items x 6 groups=90) intercepts were freed that were constrained in the model with full intercept invariance. This pattern of partial invariance of uniquenesses was used in all subsequent models with variance constraints on uniquenesses

Table 7

Patterns of Gender x Age Differences on Big-Five Latent Mean Factors (see model MAG7p in Table 6)

	Latent Means: 6 Age-Gender Groups						T-test of Statistical Significance				
	Young		Middle		Old		Linear	Quad	Sex	Sex-by	Sex-by
	M	F	M	F	M	F	Age	Age		L-Age	Q-Age
Agree	-.26	.14	-.24	.13	-.08	.31	4.67	4.37	13.99	.14	7.09
Consc	-.21	-.07	.36	.40	-.18	-.30	-2.46	-9.40	0.73	3.42	-1.98
Extra	.14	.66	-.20	.13	-.46	-.27	-16.73	-5.07	12.21	4.47	3.54
Neur	-.18	.57	-.25	.44	-.58	.00	-10.29	-7.33	18.23	1.95	11.76
Open	.60	.17	.16	-.24	-.04	-.65	-13.21	-5.85	-12.23	2.00	-8.23

Note. A=Agreeableness, C= Conscientiousness, E= Extraversion, N= Neuroticism, O= Openness. Age was divided into three categories. Presented are latent means from selected models with intercepts invariant (or partly invariant) for six groups (2 sex X 3 age groups).

Table 8

Age and Gender Effects in Big-Five Factors: Hybrid MIMIC-Multiple Group Based on Model MAG7p (See Table 6)

CHI	df	CFI	TLI	RMSEA	Description
ESEM Models With Linear (L) and Quadratic (Q) MIMIC Age Effects					
2672	752	.937	.932	.033	MIMIC-MAG0: All MIMIC L & Q Age effects = 0 (MIMIC Null)
1794	572	.960	.943	.030	MIMIC-MAGS: MIMIC L & Q Age on all 15 indicators (MIMIC Saturated)
2299	692	.947	.938	.032	MIMIC-MAG1: MIMIC L & Q Age on all 5 Latent Means
2145	674	.951	.942	.031	MIMIC-MAG2: MIMIC-MAG1with items' intercepts partial invariance (L-Age on 3 items freed across 6 groups) [a]
2257	689	.948	.939	.031	MIMIC-MAG3: MIMIC-MAG2 with items' intercepts partial invariance over 6 groups [b]
2182	686	.951	.942	.030	MIMIC-MAG4: MIMIC-MAG3 with items' intercepts partial invariance in 5 of 6 groups [b]
2218	716	.950	.944	.030	MIMIC-MAG5: MIMIC-MAG4 with all Q-Age effects constrained to be zero [c]
2365	731	.949	.944	.030	MIMIC-MAG6: MIMIC-MAG5 with MIMIC Age effects invariant over gender within age groups [d]
2294	732	.948	.943	.030	MIMIC-MAG7: MIMIC-MAG6 with MIMIC Age effects invariant over gender within age groups (11 retained) [e]
2307	738	.948	.943	.030	MIMIC-MAG8: MIMIC-MAG7 with 16 of 30 small MIMIC Age effects invariant over gender within age groups [f]

Note. CHI= chi-square; df=degrees of freedom ratio; CFI= Comparative fit index; TLI=Tucker-Lewis Index; RMSEA= Root Mean Square Error of Approximation. CUs = a priori correlated uniquenesses based on the negatively worded items. L-Age & Q-Age = linear and quadratic components of age in the MIMIC model. All analyses were weighted by the appropriate weighting factor and based on a complex design option to account for nesting within household.

a Paths from MIMIC L-Age to 3 of 15 FFA items freed (these were one previously identified in the all-MIMIC model). B. partial invariance constraints that were freely estimated across off six groups (in MIMIC-MAG2) were constrained to be equal across sex groups or in five of the six groups. D. Q-Age effects constrained to be zero across all six age-gender groups. Effects on MIMIC L-Age were constrained to be equal across responses by men and women in the same age group. f Small L-Age effects (based on contributions to goodness of fit) on latent means are constrained to zero.

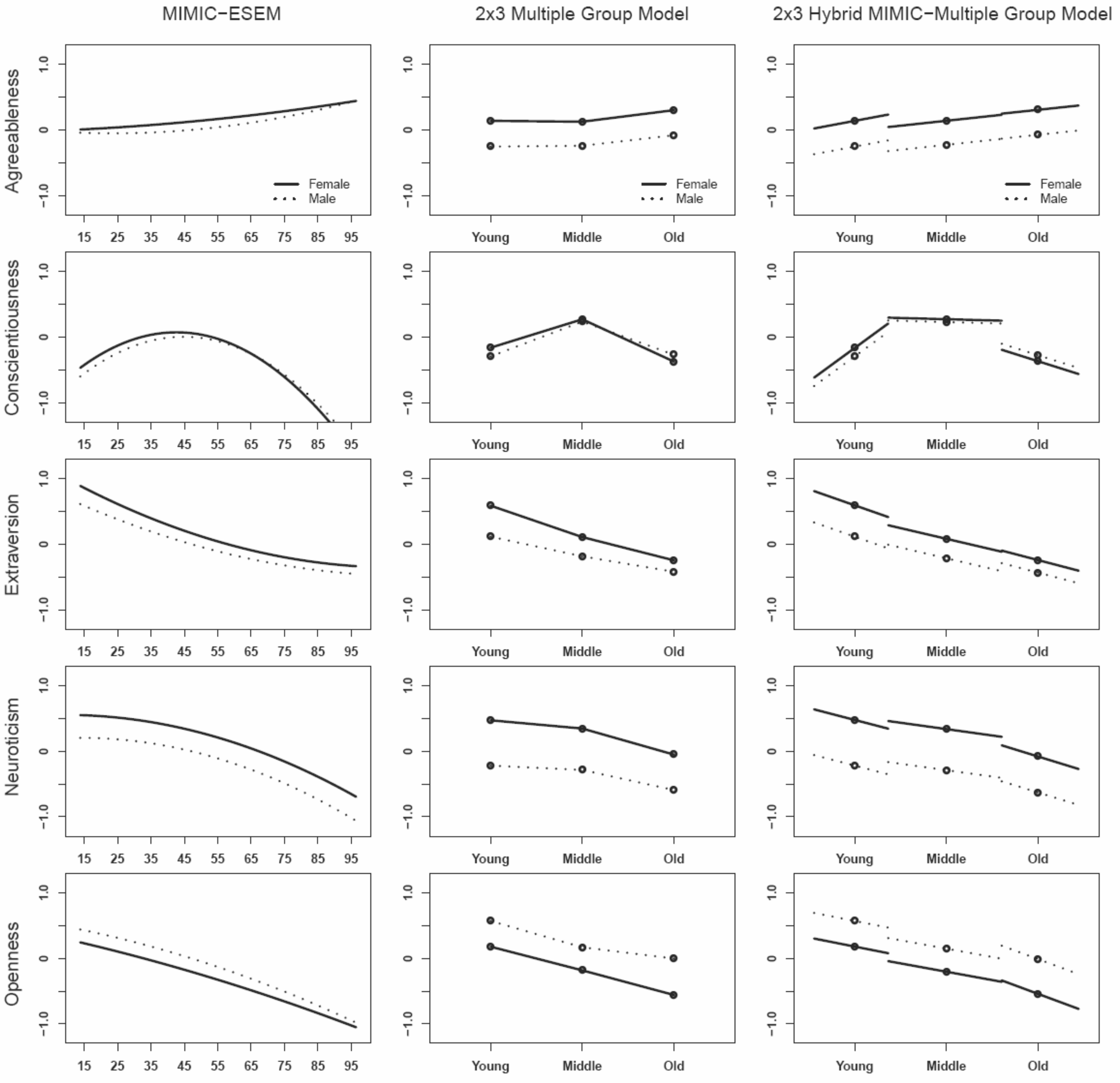

Figure 1. Alternative ESEM models of the effects of Age, Gender, and their Interaction on each of the big-five personality factors. Graphs on the left are based on the MIMIC model, those in the middle on a multiple-groups model and those on the left on the hybrid MIMIC-Multiple-Group model. Values in all the graphs were scaled to represent standardized variables (Mean = 0, SD = 1).

Supplemental Material (to go onto APA Website, hotlinked to published article)

1. Detailed Description of the Five-Factor Approach (FFA) Measures of Personality Uses in the British Household Panel Survey (BHPS)

2. The Exploratory Structural Equation Modelling (ESEM) Approach

3. MIMIC/Multiple-group Hybrid Model of Age Effects.

4. ESEM-Within-CFA Models (ES-W-C)**: An Extension of ESEM.**

5. References in Supplemental Materials (not in main article)

Detailed Description of the Five-Factor Approach (FFA) Measures of Personality Uses in the British Household Panel Survey (BHPS)

In documentation provided in the BHPS Technical Manual, (Taylor, et al., 2009, pp. a3-21 to a3-22) the FFA factors are described as follows:

> *Extraversion refers to individual differences in sociability, gregariousness, level of activity, and the experience of positive affect. Agreeableness refers to individual differences in altruistic behavior, trust, warmth, and kindness. Conscientiousness refers to individual differences in self-control, task-orientation, and rule-abiding. Neuroticism refers to individual differences in the susceptibility to distress and the experience of negative emotions such as anxiety, anger, and depression. Finally, Openness to Experience refers to individual differences in the propensity for originality, creativity, and the acceptance of new ideas. The general agreement on the Big Five provides a standardized language for describing personality differences at the broadest levels and has facilitated the accumulation of knowledge concerning how personality traits are related to a broad range of life outcomes. Personality traits tend to be assessed using long questionnaires. However, recent scale-development studies have indicated that the Big Five traits can be reliably assessed with a small number of items (e.g., Gosling et al., 2003). […] Large scale nationally represented data are crucial for establishing that personality traits are in fact essential psychological constructs. Including personality in the BHPS will make it one of the best datasets in the world for study how personality traits are linked with real-world choices and reactions over time.*

Wording of the 15 items used to Measure the big-five personality items
I see myself as someone who . . .

(A1R) 1 Is sometimes rude to others (reverse-scored).
(A2) 6 Has a forgiving nature.
(A3) 11 Is considerate and kind to almost everyone.

(C1) 2 Does a thorough job.
(C2R) 7 Tends to be lazy (reverse-scored).
(C3) 12 Does things efficiently.

(E1) 3 Is talkative.
(E2) 8 Is outgoing, sociable.
(E3R) 13 Is reserved (reverse-scored).

(N1) 4 Worries a lot.
(N2) 9 Gets nervous easily.
(N3R) 14 Is relaxed, handles stress well (reverse-scored).

(O1) 5 Is original, comes up with new ideas.

(O2)	10	Values artistic, aesthetic experiences.
(O3)	15	Has an active imagination

For each item, there is a variable label (in parentheses) that identifies the big-five factor the item was designed to measure and whether the item was reverse-scores (indicated by the suffix R). The number outside parentheses refers to the ordering of item in the actual British Household Panel Study.

The Exploratory Structural Equation Modelling (ESEM) Approach

In the ESEM model (Asparouhov & Muthén, 2009; Marsh, Muthén, et al., 2009), there are p dependent variables Y = (Y1, ..., Yp) and q independent variables X = (X1, ...,Xq) and m latent variables $\eta = (\eta_1, ..., \eta_m)$ under the standard assumptions that the ε and ζ are normally distributed residuals with mean 0 and variance covariance matrix θ and ψ respectively. Λ is a factor loading matrix, whilst B and Γ are matrices of regression coefficients relating latent variables to each other.

$$Y = \nu + \Lambda\eta + \mathrm{K}\mathrm{X} + \varepsilon \qquad (1)$$

$$\eta = \alpha + \mathrm{B}\eta + \Gamma\mathrm{X} + \xi \qquad (2)$$

Although all parameters can be identified with the maximum likelihood estimation method (ML), the model is generally not identified unless additional constraints are imposed. As in CFA analyses, the two typical approaches are to identify the metric of the latent variable by either fixing the variance of the latent variable to be 1.0 or by fixing one of the factor loadings for each factor typically to be 1.0.

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. In particular, when more than one factor is posited (m > 1.0), further constrains are required to achieve an identified solution. To resolve this problem, consider any m x m square matrix (m = number of factors), a square matrix that we refer to as H. In this (mxm) square matrix H one can replace the η vector by H η in the ESEM model (1-2) which will also alter the parameters in the model as well; Λ to ΛH^{-1}, the α vector H α, the Γ matrix to H Γ, the B matrix to HBH^{-1} and the Ψ matrix to $H\Psi H^{T}$. Since H has m^2 elements, the ESEM model has a total of m^2 indeterminacies that must be resolved. Two variations of this model are considered; one where factors are orthogonal so that the factor variance-covariance matrix (Ψ) is an identity matrix, and an oblique model where Ψ is an unrestricted correlation matrix (i.e., all correlations and residual correlations between the latent variables are estimated as free parameters). This model can also be extended to include a structured variance-covariance matrix (Ψ).

For an orthogonal matrix H (i.e., a square mxm matrix H such that $HH^T = I$), one can replace the η vector by H η and obtain an equivalent model in which the parameters are changed. EFA can resolve this non-identification problem by minimizing $f(\Lambda^*) = f(\Lambda H{-}1)$, where f is a function called the rotation criteria or simplicity function (Asparouhov & Muthén, 2009; Jennrich & Sampson, 1966), typically such that among all equivalent Λ parameters the simplest solution is obtained. There are a total of $m(m-1)/2$ constraints in addition to $m(m + 1)/2$ constraints that are directly imposed on the Ψ matrix for a total of m^2 constraints needed to identify the model. The identification for the oblique model is developed similarly such that a total of m^2 constraints needed to identify the model are imposed. Although the requirement for m^2 constraints is only a necessary condition and in some cases it may be insufficient, in most cases the model is identified if and only if the Fisher information matrix is not singular (Silvey, 1970). This method can be used in the ESEM framework as well (Asparouhov & Muthén, 2009; also see Hayashi & Marcoulides, 2006).

The estimation of the ESEM model consists of several steps (Asparouhov & Muthén, 2009). Initially a SEM model is estimated using the ML estimator. The factor variance covariance matrix is specified as an identity matrix (ψ = I), giving $m(m + 1)/2$ restrictions. The EFA loading matrix (Λ), has all entries above the main diagonal (i.e., for the first m rows and column in the upper right hand corner of factor loading matrix, Λ), fixed to 0, providing remaining $m(m - 1)/2$ identifying restrictions. This initial, unrotated model provides starting values that can be subsequently rotated into an EFA model with m factors. The asymptotic distribution of all parameter estimates in this starting value model is also obtained. Then the ESEM variance covariance matrix is computed (based only on $\Lambda\Lambda^T + \theta$ and ignoring the remaining part of the model).

The correlation matrix is also computed and, using the delta method (Asparouhov & Muthén, 2009), the asymptotic distribution of the correlation matrix and the standardization factors are obtained. In addition, again using the delta method, the joint asymptotic distribution of the correlation matrix, standardization factors and all remaining parameters in the model are computed and used to obtain the standardized rotated solution based on the correlation matrix and its asymptotic distribution (Asparouhov & Muthén, 2009). This method is also extended to provide the asymptotic covariance of the standardized rotated solution, standardized unrotated solution, standardization factors, and all other parameters in the model. This asymptotic covariance is then used to compute the asymptotic distribution of the optimal rotation matrix H and all unrotated parameters which is then used to compute the rotated solution for the model and its asymptotic variance covariance.

In Mplus multiple random starting values are used in the estimation process to protect against non-convergence and local minimums in the rotation algorithms. Although a wide variety of orthogonal and oblique rotation procedures are available, leading authorities on this topic (e.g., Asparouhov & Muthén, 2009; Browne, 2001; Jennrich, 2006) have recommended Geomin rotation, but made it clear that the researchers should explore alternative solutions with different rotation strategies. In the context of the present investigation, geomin ration had a desirable theoretical and statistical rationales in that it was developed specifically to better represent simple structure as conceived by Thurstone (1947) which is very different to how it has sometimes been interpreted and clearly inconsistent with the ICM-CFA model. Geomin rotations also incorporate a complexity parameter consistent with Thrustone's original proposal. As operationalized in Mplus, this complexity parameter (ε) takes on small positive value that increases with the number of factors (Browne, 2001; Asparouhov & Muthén, 2009). In the present investigation we found that increasing the ε altered the balance between the sizes of cross-loadings and factor correlations. As we were especially concerned with the sizes of factor correlations, we set the epsilon at a rather high value (.5) that resulted in somewhat lower factor correlations and somewhat higher cross-loadings. Nevertheless, consistent with recommendations, we explored a number of different rotations in preliminary analyses. There did not seem to be substantial differences results based on the various rotations, consistent with suggestions by Asparouhov & Muthén (2009) who concluded that "*In most ESEM applications the choice of the rotation criterion will have little or no effect on the rotated parameter estimates"* (p. 428)**.** Although we had a clear basis for using the geomin rotation, we are not suggesting that this will always – or even generally – be the best rotation in other studies. Quite the contrary, following recommendations based on Asparouhov and Muthén (2009), Browne (2001) and others – as well as our own personal experience, we suggest that applied researchers should evaluate the theoretical and mathematical rationales for difference rotations, experiment with a number of different rotations and complexity parameters, and chose the one that is most appropriate for their specific application. We also note that this is clearly an area where more research – using both simulation and read data – is needed.

With ESEM models it is possible to constrain the loadings to be equal across two or more sets of EFA blocks in which the different blocks represent multiple discrete groups or multiple occasions for the same group. This is accomplished by first estimating an unrotated solution with all loadings constrained to be equal across the groups or over time. If the starting solutions in the rotation algorithm are the same, and no loading standardizing is used, the optimal rotation matrix will be the same as well as the subsequent rotated solutions. Thus obtaining a model with invariant rotated Λ^* amounts to simply estimating a model with invariant unrotated Λ, a standard task in maximum likelihood estimation.

For an oblique rotation it is also possible to test the invariance of the factor variance-covariance matrix (Ψ) matrix across the groups. To obtain non-invariant Ψs an unrotated solution with $\Psi = I$ is specified in the first group and an unrestricted Ψ is specified in all other groups. Note that this unrestricted specification means that Ψ is not a correlation matrix as factor variances are freely estimated. It is not possible in the ESEM framework to estimate a model where in the subsequent groups the Ψ matrix is an unrestricted correlation matrix, because even if the factor variances are constrained to be 1 in the unrotated solution, they will not be 1 in the rotated solution. However, it is possible to estimate an unrestricted Ψ in all but the first group and after the rotation the rotated Ψ can be constrained to be invariant or varying across groups. Similarly, when the rotated and unrotated loadings are invariant across groups, it is possible to test the invariance of the factor intercept and the structural regression coefficients. These coefficients can also be invariant or varying across groups simply by estimating the invariant or group-varying unrotated model. However, in this framework only full invariance can be tested in relation to parameters in Ψ and Λ in that it is not possible to have measurement invariance for one EFA factor but not for the other EFA factors. Similar restrictions apply to the factor variance covariance, intercepts and regression coefficients, although it is possible to have partial invariance in the ε matrix of residuals. (It is however, possible to have different blocks of ESEM factors such that invariance constraints are imposed in one block, but not the other). Furthermore, if the ESEM model contains both EFA factors and CFA factors, then all of the typical strategies for the SEM factors can be pursued with the CFA factors.

MIMIC/Multiple-group Hybrid Model of Age Effects.

Studies of age differences in FFA factors typically suffer the same methodological shortcomings as those of gender differences. However, the tests of invariance become even more complex in that age is a continuous variable rather that a natural categorical variable with a few discrete groups (like gender). Valid interpretations of age differences assume that that there is measurement invariance across all possible intervals of the age continuum under consideration. Multiple-group tests of invariance for continous variables require that the continuous variables be divided into a relatively small number of groups, but there are inherent dangers in transforming continuous variables into categorical variables (MacCallum, Zhang, Preacher & Rucker, 2002). Within the CFA literature, there are traditionally two approaches to this problem, each with counter-balancing strengths and limitations: The MIMIC and the multiple-group approaches (e.g., Marsh, Tracey & Craven, 2006). The MIMIC model regresses the latent variables (the FFA factors) onto other variables (continuous, like age, or categorical, like gender). However, there are important limitations in the capability of the MIMIC model to evaluate invariance assumptions: Only the invariance of item intercepts and factor means can be tested. In the multiple group approach, it is possible to pursue the more rigorous tests of invariance presented in Table 1. However, for continuous variables, these tests require researchers to transform continuous variables into a relatively small number of categories that constitute the multiple groups. Marsh, Tracey and Craven (2006) proposed a hybrid approach involving an integration of interpretations based on both MIMIC and multiple group approaches. The multiple group approach is used to test invariance assumptions that cannot easily be tested with the MIMIC approach and the MIMIC approach is used to infer differences in relation to a score continuum rather than discrete groups. So long as the two approaches converge to similar interpretations, there is support for the construct validity of these interpretations (e.g., Marsh, Martin & Hau, 2006).

In the present investigation, we extend this hybrid approach in two important ways. First, following Marsh, Lüdtke et al. (2010), we demonstrate how this hybrid approach can be translated into the ESEM framework so that researchers do not have to rely on potentially problematic CFA models. Second, we demonstrate how the MIMIC and multiple group approaches can both be incorporated into a single ESEM model. This means that researchers do not have to rely on the subjective integration of results from two entirely different models and do not have to rely on the fortuitous combination of results from the two models to justify interpretations of their results. Here we extend the logic of the hybrid approach by actually adding the MIMIC age (linear and quadratic) variables to the multiple group model (based on sex-age groups). This allows us to evaluate more formally whether information in the continuous age effects are lost by forming categories and, if so, to estimate the combined age effects due to both representations of age.

The interaction between gender and age is substantively important to interpretations of both gender and age effects. However, the introduction of interactions into latent variable models presents a whole host of new complications. Particularly relevant to the present investigation, invariance assumptions apply not only to the distinct gender groups and to age groups (or regions along the age continuum), but also to groups (or regions) defined by all possible combinations of these two variables. Although tests of invariance have rarely been applied to the interaction of two variables, we illustrate how the use of ESEM to the hybrid integration of multiple-group and MIMIC models can be extended to include interactions between variables.

ESEM-Within-CFA Models (ES-W-C)**: An Extension of ESEM.**

The ESEM approach is very flexible, but there are still some tests that are easily done with CFA/SEM models that cannot easily be accomplished with ESEM models as currently operationalized in Mplus (e.g., constraints on group specific correlations among factors, tests of higher-order factor models, fully latent latent curve models, factor mixture models, etc.). Of particular relevance to the present investigation, applied researchers cannot easily place constraints on latent means estimated in multiple groups models to test linear and non-linear effects based on a single grouping variable such as age or interaction effects between two grouping variables such as age-by-gender interactions. Here we propose an extension of the ESEM approach to address this limitation that makes the ESEM approach much more flexible – what we refer to as ESEM-Within-CFA Models (ES-W-C).

The ES-W-C approach introduced here is based in large part on an EFA-Within-CFA proposal by Jöreskog (1969; also see Muthén & Muthén, 2009, slides 133-146). EFA-Within-CFA provided SEs for EFA parameter estimates and greater flexibility in specifying factor structures; potentially important limitations in EFA at that time. However, SEs for EFA solutions, the inclusion CUs, and a variety of goodness of fit measures are now available through standard statistical packages such as the ESEM routine in Mplus (also see Dolan et al., 2009). Nevertheless, for the aforementioned cases and when preliminary analyses show that ESEM solutions are superior to CFA solutions, ES-W-C may represent an interesting complementary alternative to ESEM.

As the original EFA-within-CFA model must contain the same number of restrictions as the EFA– m^2 (where m=number of factors) – it was proposed that researchers look for referent indicators that have near-zero loadings on all but the latent constructs that they are designed to measure (from initial runs of EFA solutions). The cross-loadings for these indicators were then constrained to be zero in a corresponding CFA model where each of the remaining loadings on all factors are freely estimated and the factors variances are constrained to be 1. Under appropriate conditions, the resulting CFA parameter estimates will approximate the corresponding EFA model. In our ES-W-C approach, we offer an amendment to this approach:

- In preliminary analyses, compare ESEM and CFA models in relation to goodness of fit and parameter estimates. If the ESEM solution is not clearly superior, ESEM or ES-W-C models should not be pursued.
- If preliminary ESEM models are better than CFA models, a complete ESEM analysis should be done to identify the best model in relation to goodness of fit and substantive interpretations.
- If there is a need to conduct an additional analysis that cannot be easily implemented within the ESEM framework, then all parameter estimates from the final ESEM solution should be used as starting values to estimate the ES-W-C model.
- We now need to add a total of m^2 constraints (where m=number of factors) to the ES-W-C model so that it is identified. This is usually done by fixing selected parameter estimates to the values obtained from the ESEM solution. One convenient way to do this is to:
    - Fix the m factor variances (by default these parameters are fixed to be 1.0 in a single-group ESEM solution or for the first group of a multiple group solution).
    - Select an “anchor“ item (or referent indicators) for each factor that has a large loading for that factor it is designed to measure and small cross-loadings on the other factors. Then, fix these small cross loadings to their values from the ESEM solution. This allows for a higher level of convergence with the ESEM solution than the classical method of fixing these cross-loadigns to zero. (Because the ESEM parameter estimates have already been included, this merely means fixing these parameter estimates so that they are not freely estimated).
    - For all other parameter estimates, the pattern of fixed and free estimates should be the same as in the selected ESEM solution (i.e., if the parameter is free in the ESEM solution it is free in the ES-W-C solution; if the parameter is fixed or constrained in the ESEM solution it should be the same in the ES-W-C solution).
    - It should be noted that the mean structure from the ES-W-C solution can be identified as in a standard CFA model (while using the ESEM start values when possible). This is usually done by freely estimating all items intercepts and constraining all factor means to zero (or the factor means from the first group of a multiple group solution).

The ES-W-C solution will have the same df and, within rounding error, the same chi-square value, the same goodness of fit statistics, and – most importantly – the same parameter estimates. In this sense, it is equivalent to the ESEM solution. The ES-W-C approach is stronger then the EFA-Within-CFA approach in that the parameter estimates are the same as the corresponding ESEM approach rather than approximations, particularly when anchor items with zero factor loadings do not exist. Furthermore, the ESEM approach used to generate the starting parameter estimates for the ES-W-C approach is more flexible than the traditional EFA approach used in the EFA-Within-CFA strategy. Importantly, the applied researcher has more flexibility in terms of how to constrain or further modify the ES-W-C model (as it is a true CFA model) than with the ESEM model upon which it is based.

In the present investigation, for example, we added constraints on the estimated latent means to represent linear and quadratic effects of age, gender, and age-by-gender interactions – and tests of statistical significance of these effects (see appendix for Mplus codes). Because these constraints are merely a re-parameterization of the observed means, they did not alter the goodness of fit and other parameter estimates. However, it also possible to add further limitations or estimate additional parameters that do change the df and thus influence other parameter estimates (e.g., freeing or constraining parameters in specific groups). Particularly if the changes are substantial, it is important to evaluate critically the new parameter estimates in relation to substantive interpretations as well as to assess changes in goodness of fit, but this is also the case for any ESEM or CFA solution. Although the ES-W-C approach is very general, it is particularly useful in defining contrasts on latent means (like those used in traditional ANOVA approaches) and testing the invariance of parameters linked to these constraints that cannot normally be evaluated in the ESEM approach. Nevertheless, as this approach has not been widely used in applied research, caution is needed in its application when the ES-W-C model differs substantially from the ESEM model upon which it was based.

Model constraints in ES-W-C used to define contrasts to test main and interaction effects of sex and age are as following:

Mplus Syntax Converting ESEM solution Into CFA solution (with unit Weights)

```
      TITLE:
        DATA: FILE IS bhpsindrespX4.DAT;

         VARIABLE: NAMES ARE
         b5a1r b5a2 b5a3 b5c1 b5c2r b5c3 b5e1 b5e2 b5e3r b5n1 b5n2 b5n3r
          b5o1 b5o2 b5o3 fsb5n fsb5a fsb5e fsb5c fsb5o ohid a3s2 oXRWTUK2;
  ! b5a1r-b5o3 are the 15 items that define the FFA factors;
  ! ohid = household cluster id variable;
  ! a3s2= group identifiers for the 6 (3 age x 2 sex) groups;
  ! oXRWTUK2 = weighting variable supplied with the data;

         usevariables   b5a1r-b5o3 ohid;
         weight is oXRWTUK2; ! sampling weight
         missing are all (-9.99); ! little missing, default=FIML
          cluster is ohid; ! complex design cluster by household
         GROUPING IS a3s2 (11=A1S1, 12=A1S2,21=A2S1,22=A2S2,31=A3S1,32=A3S2);
            ! grouping variable that defines the 6 groups (3 age x 2 gender)

        ANALYSIS: type=complex; ESTIMATOR=MLR;
! syntax for the first group (A1S1) in which starting values (*) and fixed values (@)
! are copied from the ESEM results. Numbers in parentheses are used to make constrained
! parameter estimates to be invariant across two or more groups (i.e. parameters with the
! same number in parentheses are constrained to equality).

! Model in the first group.
  Model A1S1:
! Factor loadings of all items on Factor 1. Thus item B5A1R loads on factor F1.
! The factor loading is freely estimated (as indicated the by *),
!          but has a starting value of 0.603
! The factor loadings of the anchor items are fixed as indicated by the "@"
!          instead of the * (e.g. the factor loading of item B5C3 on F1 is fixed at 0.198)
  F1 BY  B5A1R  * 0.603;
  F1 BY  B5A2 *   0.729;
  F1 BY  B5A3   * 0.869;
  F1 BY  B5C1  * -0.073;
  F1 BY  B5C2R  * 0.010;
  F1 BY  B5C3   @ 0.198;
  F1 BY  B5E1   @ 0.018;
  F1 BY  B5E2   * 0.279;
  F1 BY  B5E3R * -0.239;
  F1 BY  B5N1  @ -0.043;
  F1 BY  B5N2   * 0.144;
  F1 BY  B5N3R * -0.310;
  F1 BY  B5O1  * -0.093;
  F1 BY  B5O2   * 0.119;
  F1 BY  B5O3   @ 0.130;
! A similar section should be added for the remaining factors
 ● ● ● ● ● ● ● ●
! Factor covariances;
  F2 with F1 *  0.265  ;
  F3 with F1 *  0.134  ;
```

```
  F3 with F2 *  0.188  ;
  F4 with F1 * -0.067  ;
  F4 with F2 * -0.134  ;
  F4 with F3 * -0.206  ;
  F5 with F1 *  0.236  ;
  F5 with F2 *  0.214  ;
  F5 with F3 *  0.275  ;
  F5 with F4 *  0.072  ;
! Correlated uniquenesses;
  B5A1R WITH B5C2R *   0.345 (1) ;
  B5A1R WITH B5E3R *   0.035 (2);
  B5A1R WITH B5N3R *   0.108 (3) ;
  B5C2R WITH B5E3R *   0.162 (4);
  B5C2R WITH B5N3R *   0.161 (5);
  B5E3R WITH B5N3R *   0.243 (6);
! the numbers in parentheses are used to label a particular parameter estimate so that it can be
!       referenced in future constraints (e.g., constrained to be invariant over groups).
! Items Intercepts;
  [ b5a1r *  5.539 ] (31)   ;
  [ B5A2 *   5.014 ] (32)  ;
  [ B5A3 *   5.159 ] (33)  ;
  [ B5C1 *   5.161 ] (34)  ;
  [ B5C2R *  4.389 ] (135)  ;
  [ B5C3 *   5.104 ] (36)  ;
  [ B5E1 *   4.681 ] (37)  ;
  [ B5E2 *   5.183 ] (138)  ;
  [ B5E3R *  4.214 ] (39)  ;
  [ B5N1 *   3.387 ] (140)  ;
  [ B5N2 *   3.476 ] (141)  ;
  [ B5N3R *  3.477 ] (42)  ;
  [ B5O1 *   4.778 ] (43)  ;
  [ B5O2 *   4.296 ] (144)  ;
  [ B5O3 *   5.307 ] (45)  ;


! Factor Means (constrained to be zero in the first group);
  [F1 @   0.000] ;
  [F2 @   0.000] ;
  [F3 @   0.000] ;
  [F4 @   0.000] ;
  [F5 @   0.000] ;
! Factor variances (constrained to be 1 in the first group);
      F1     @ 1.0 ;
      F2     @ 1.0 ;
      F3     @ 1.0 ;
      F4     @ 1.0 ;
      F5     @ 1.0 ;
  !Residual Variances;
      B5A1R *  1.596 (51) ;
      B5A2 *   1.515 (52) ;
      B5A3 *   0.480 (53) ;
      B5C1 *   1.415 (54) ;
      B5C2R *  2.159 (55) ;
      B5C3 *   0.346 (56) ;
      B5E1 *   0.967 (57)  ;
      B5E2 *   1.115 (58)  ;
      B5E3R *  1.715 (59)  ;
      B5N1 *   1.345 (60)  ;
      B5N2 *   1.492 (61)  ;
      B5N3R *  1.345 (62)  ;
      B5O1 *   0.927 (63)  ;
      B5O2 *   1.664 (64)  ;
      B5O3 *   1.121 (65)  ;

! New variables are defined as the means of the latent factors.
! A1S1_m1 = mean of F1 (first factor) in group A1S1 (Age = 1, s = 1; young males)
         [f1](A1S1_m1);  [f2](A1S1_m2); [f3](A1S1_m3); [f4](A1S1_m4); [f5](A1S1_m5);
! The fixed and free values for the remaining 5 groups are specified according to
! the pattern of fixed and free values and parameter estimates in the selected ESEM solution;
 ● ● ● ● ● ● ● ●
! Latent means for the same factor across the six age-gender groups are constrained
! to sum to 0 (for purposes of identification);
  MODEL CONSTRAINT: !factor 1
        0 = A1S1_m1 + A2S1_m1 + A3S1_m1 + A1S2_m1 +  A2S2_m1 + A3S2_m1;
        0 = A1S1_m2 + A2S1_m2 + A3S1_m2 + A1S2_m2 +  A2S2_m2 + A3S2_m2;
        0 = A1S1_m3 + A2S1_m3 + A3S1_m3 + A1S2_m3 +  A2S2_m3 + A3S2_m3;
        0 = A1S1_m4 + A2S1_m4 + A3S1_m4 + A1S2_m4 +  A2S2_m4 + A3S2_m4;
        0 = A1S1_m5 + A2S1_m5 + A3S1_m5 + A1S2_m5 +  A2S2_m5 + A3S2_m5;

! Using traditional contrasts as in a typical linear model, groups means are used to define the
!      main effects of gender (male vs. female), age (linear and quadratic) and their interaction.
! Here are the contrasts for factor 1 (F1) based on the six age-by-sex latent means
! (A1S1 - A3S2) for the first factor (_m1)
        NEW(MSEXF1);
            MSEXF1 =  (+1* A1S1_m1) +  (+1* A2S1_m1) + (+1* A3S1_m1) +
                      (-1* A1S2_m1) + (-1* A2S2_m1) + (-1* A3S2_m1);
        NEW(lAGEF1);
             LAGEF1=  (-1.0* A1S1_m1) + (0* A1S2_m1) + (+1.0* A3S1_m1) +
```

```
              (-1.0* A1S2_m1) + (0* A2S2_m1) + (+1.0* A3S2_m1);
NEW(QAGEf1);
    QAGEf1=(1.0* A1S1_m1) + (-2.0* A1S2_m1) + (1.0* A3S1_m1) +
           (1.0* A1S1_m1) + (-2.0* A2S2_m1) + (1.0* A3S2_m1);
NEW(SXAGLF1);
   SXAGLF1=  (-1.0* A1S1_m1) + (0* A1S2_m1) + (+1.0* A3S1_m1) +
             (+1.0* A1S2_m1) + (0* A2S2_m1) + (-1.0* A3S2_m1);
NEW(SXAGQF1);
     SXAGQF1=(+1.0* A1S1_m1) + (-2.0* A2S2_m1) + (+1.0* A3S1_m1) +
             (-1.0* A1S1_m1) + (2.0* A2S2_m1) + (-1.0* A3S2_m1);
```

● ● ● ● ● ● ● ●

```
! The same pattern of constraints is applied to each of the FFA factors (F2-F5) for
! latent means for the remaining five factors (_m2, _m3, _m4, _m5).
```

References in Supplemental Materials (not in main article)

Browne, M. W. (2000). Cross-validation methods. Journal of Mathematical Psychology, 44, 108–132.

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

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

Hayashi, K., & Marcoulides, G. A. (2006). Examining identification issues in factor analysis. Structural Equation Modeling, 13, 631–645.

Jennrich, R. I. (2006). Rotation to Simple Loadings Using Component Loss Functions: The Oblique Case.Psychometrika, 71, 173-191.

Jennrich, R. I., & Sampson, P. F. (1966). Rotation to simple loadings. Psychometrika, 31, 313–323.

Jöreskog, K. G. (1969). A general approach to confirmatory maximun likelihood factor analysis. *Psychometrika*, *34*, 183-202.

Silvey, S. D. (1970). Statistical inference. Harmondsworth, UK: Penguin.

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