Chapter 10.

Exploratory Structural Equation Modeling

Alexandre J. S. Morin

University of Western Sydney, Australia & University of Sherbrooke, Canada

Herbert W. Marsh

University of Western Sydney, Australia, University of Oxford, UK, & King Saud University, Saudi Arabia

and

Benjamin Nagengast

University of Tübingen, Germany

To appear in
Hancock, G. R., & Mueller, R. O. (Eds.). (2013). Structural equation modeling: A second course (2nd ed.). Charlotte, NC: Information Age Publishing, Inc.

Acknowledgements

The authors would like to thank Tihomir Asparouhov and Bengt Muthén for helpful comments at earlier stages of this research, and each of the co-authors of ESEM studies cited in this chapter (Tihomir Asparouhov, Rhonda Craven, Linda Hamilton, Gregory Arief Liem, Oliver Lüdtke, Christophe Maïano, Andrew Martin, Bengt Muthén, Roberto Parada, Alexander Robitzsch, and Ulrich Trautwein). This research was supported in part by a grant to the second author from the UK Economic and Social Research Council.

Exploratory Structural Equation Modeling

In a seminal publication, Cohen (1968) presented multiple regression as a generic data-analytic system for quantitative dependent variables (outcomes) encompassing classical analyses of variance and covariance, interactive effects, and predictive non-linearity among quantitative and qualitative predictors. However, multiple regression and most of the General Linear Model procedures were later shown to represent special cases of the even more encompassing framework of canonical correlation analysis, allowing for the inclusion of multiple outcomes within the same model (Knapp, 1978). Similarly, Structural Equation Modeling (SEM) was proposed as an even more flexible framework (Bagozzi, Fornell, & Larker, 1981; Fan, 1997; Graham, 2008), covering any relation that could be studied with canonical correlation analysis, but, also allowing for the simultaneous estimation of chains of direct and indirect effects (i.e. path analysis) based on latent variables that implicitly correct the estimated relations for measurement error (i.e., confirmatory factor analysis [CFA]). Then, Muthén (2002; also see Skrondal & Rabe-Hesketh, 2004) incorporated all of these methods into an even more generic framework (generalized SEM [GSEM]), allowing for the estimation of relations between any type of quantitative or qualitative observed and latent variables. Although exploratory factor analyses (EFAs) have been around for more than a century (Spearman, 1904) and represent an important precursor of CFAs (Cudeck & MacCallum, 2007), and thus of SEM and GSEM, it was only until recently kept out of these generalized frameworks.

It is thus not surprising that EFA is now seen as less useful, and even outdated, compared to the methodological advances associated with CFA/SEM (e.g., estimation of structural relations among latent constructs adjusted for measurement error, higher order factor models, method factors, correlated uniquenesses, goodness of fit assessment, latent mean structures, differential item functioning/measurement invariance, and latent curve

modeling). This perception is reinforced by the erroneous semantically-based assumption that EFA is strictly an *exploratory* method that should only be used when the researcher has no a priori assumption regarding factor structure and that *confirmatory* methods are better in studies based on a priori hypotheses regarding factor structure. This assumption still serves to camouflage the fact that the critical difference between EFA and CFA is that all cross loadings are freely estimated in EFA. Due to this free estimation of all cross loadings, EFA is clearly more naturally suited to exploration than CFA. However, statistically, nothing precludes the use of EFA for confirmatory purposes, except perhaps the fact that most of the advances associated with CFA/SEM were not, until recently, available with EFA.

In fact, many psychological instruments with well-defined EFA structures are not supported by CFAs. A classic example is the Big-Five personality factor structure that has been consistently identified and supported by an impressive body of research relying on EFAs (see McCrae & Costa, 1997). However, CFAs have failed to replicate these findings when based on analyses of the items forming Big-Five questionnaires. On the one hand, these discrepant findings led some scholars (e.g., Vassend & Skrondal, 1997) to question the factor structure of the NEO instruments—the most widely used big-five personality instruments—as well as other big-five personality instruments. On the other hand, many researchers questioned the appropriateness of CFA for Big-Five research in principle (see Borkenau & Ostendorf, 1993; Church & Burke, 1994; McCrae, Zonderman, Costa, Bond, & Paunonen, 1996; Parker, Bagby, & Summerfeldt, 1993). For instance, McCrae et al. (1996) concluded: "In actual analyses of personality data […] 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" (p. 568). Marsh et al. (2009) presented a similar argument based on the Students' Evaluations of Educational Quality (SEEQ) instrument, whose structure also received strong support from EFAs (Marsh, 1983, 1987; Marsh & Hocevar, 1991), whereas CFA models

failed to replicate these results (e.g., Toland & De Ayala, 2005). In addressing this issue, Marsh (1991a, 1991b) showed that the reason CFA structures did not provide an adequate fit to the data was that many items had minor cross-loading on other factors.

More generally Marsh, Hau, and Grayson, 2005 (also see Marsh, 2007) made the claim that “it is almost impossible to get an acceptable fit (e.g., CFI, RNI, TLI > .90; RMSEA < .05) for even "good" multifactor rating instruments when analyses are done at the item level and there are multiple factors (e.g., 5-10), each measured with a reasonable number of items (e.g., at least 5-10/per scale) so that there are at least 50 items overall” (p. 325). Marsh (30 August 2000, SEMNET@BAMA.UA.EDU) placed this claim on SEMNET (an electronic network devoted to SEM) and invited the more than 2000 members to provide counter-examples. No one offered a published counter-example, leading him to conclude that there are probably many psychological instruments routinely used in applied research that do not even meet minimum criteria of acceptable fit according to current standards.

More recently, however, Marsh and colleagues (Marsh, Liem, Martin, Morin, & Nagengast, 2011; Marsh, Nagengast, et al, 2011; Marsh, Nagengast, & Morin, 2012; Marsh et al., 2009, 2010) suggested that the independent cluster model inherent in CFA (ICM-CFA)—in which items are required to load on one, and only one, factor, with non-target loadings constrained to be zero—could be too restrictive for many multidimensional constructs. They also noted that in many CFA applications, even when the ICM-CFA model fits well (Marsh, Liem, et al., 2011), factor correlations are likely to be inflated unless all non-target loadings are close to zero (Marsh, Lüdtke, et al., 2011). This in turn undermines the discriminant validity of the factors and results in multicollinearity in the estimation of relations with outcomes (Marsh et al., 2010). Recent simulation studies seem to support that EFAs are better at recovering true population latent correlations and that CFA-based latent correlations can be severely inflated by the presence of even a few small cross loadings erroneously fixed

to zero (Asparouhov & Muthén, 2009; Marsh, Lüdtke et al., 2011). Although there are advantages in having “pure” items that load only on a single factor, this is not a requirement of a well-defined factor structure, nor even a requirement of traditional definitions of simple structure in which non-target loadings are ideally small relative to target loadings but not required to be zero (Carroll, 1953; McDonald, 1985; Thurstone, 1947). Further, strategies that are often used to compensate for ICM-CFA models' inadequacies (e.g., parceling, ex post facto modifications such as ad hoc correlated uniquenesses) tend to be dubious, misleading, or simply wrong (Browne, 2001; Marsh et al., 2009, 2010). Why then do researchers persist with CFA models even when they have been shown to be inadequate? In fact, the recent dominance of CFAs in the literature might have generated the mistaken belief that EFAs are no longer viable or even acceptable. In addition, as we pointed out earlier, many recent advances in latent variable modeling have been reserved for CFA/SEM. Hence, failure to embrace these new and evolving methodologies could have unfortunate consequences for applied research.

Fortunately, this extreme solution is no longer necessary with the development of exploratory structural equation modeling (ESEM) by Asparouhov and Muthén (2009). More specifically, ESEM is an integration of EFA within the global Muthén (2002) GSEM framework. Thus, ESEM combines the benefits associated with EFA’s flexibility along with access to typical CFA/SEM parameters and statistical advances: standard errors; goodness of fit statistics; comparisons of competing models through tests of statistical significance and fit indices; inclusion of correlated uniquenesses; inclusion of both CFA and EFA factors based on the same, different, or overlapping sets of items; estimation of method effects; multiple indicators multiple causes models (MIMIC) models; and tests of multiple group and longitudinal invariance. Currently, ESEM is available in the commercial Mplus package (Muthén & Muthén, 2010), starting from version 5.1. Before moving on, we should

acknowledge that, although Asparouhov and Muthén (2009) were the first to manage the integration of EFA within the CFA/SEM framework, previous efforts in this direction were made by Dolan, Oort, Stoel, and Wicherts (2009; see also Hessen, Dolan, & Wicherts, 2006), who developed ways to test for multiple group invariance of EFA solutions and simultaneous rotation in multiple groups in the Mx package (Neale, Boker, Xie, & Maes, 2003). Also, standard errors for rotated EFA solutions along with goodness-of-fit statistics and model tests have been available in the CEFA program (Browne, Cudeck, Tateneni, & Mels, 2010), although without capabilities for testing measurement invariance across multiple groups.

Previous applications of ESEM

To date, ESEM applications are still few. In the first of those, Marsh et al. (2009) used ESEM to analyze over 30,000 sets of class-average responses to the SEEQ instrument, designed to evaluate students' evaluation of university teaching effectiveness. Previous research provided strong support for the EFA structure, but ICM-CFAs of SEEQ responses did not fit the data. Marsh et al. found support for their claim that ICM-CFA was inappropriate for this instrument and resulted in substantially inflated factor correlations among the nine SEEQ factors (median *r*s = .34 for ESEM and .72 for CFA). They also showed that the SEEQ ESEM factor model was reasonably invariant over time and extended the ESEM model to approximate latent growth models using linear and quadratic functions of time as a MIMIC-like predictor of the factors. Finally, they conducted an extended set of MIMIC analyses to test that the potential biases (workload/difficulty, class size, prior subject interest, expected grades) to SEEQ responses were small in size and varied systematically for different ESEM SET factors, supporting a construct validity interpretation of the relations.

Marsh et al. (2010) conducted similar multiple group and longitudinal analyses on the answers provided by a sample of over 1500 high school students to the 60-item NEO Five Factor Inventory. Previous research based on this instrument found that the a priori (i.e.,

confirmatory) EFA factor structure was well defined, but that ICM-CFA models provided a poor fit to the data. Results showed that ESEM factors were more differentiated (less correlated) than ICM-CFA factors and that the ESEM factor structure was reasonably invariant over gender and time. More recently, Marsh et al. (2012) pursued extended analyses of Big-Five data to cover a lifespan perspective using data from the nationally representative British Household Panel Study (*N*=14,021 participants aged 15-99). These authors contrasted three hypotheses describing personality evolution over the lifespan (i.e., the maturity principle, the plaster hypothesis, and the Dolce Vita hypothesis). Once again, ESEM resulted in a better fit to the data and less differentiated factors than ICM-CFA. Analyses were based on a set of multiple group, MIMIC, hybrid multiple-group MIMIC models (designed to investigate the loss of information due to the categorization of continuous variables) and introduced ESEM-within-CFA (EWC), an extension designed to further increase the flexibility of ESEM.

Marsh, Liem, et al. (2011) and Marsh, Nagengast, et al. (2011) used ESEM to conduct extensive psychometric evaluations—including multi-trait multi-method analyses of the construct validity—of the Motivation and Engagement Scale and the Adolescent Peer Relations Instrument (APRI). Interestingly, both of these studies showed that ESEM might provide a more appropriate and realistic representation of the data—with substantially deflated factor correlations, even when the comparative ICM-CFA model fits the data reasonably well in the first place. In particular, Marsh, Nagengast, et al. conducted a multitrait-multimethod analysis of the APRI and showed that latent ESEM factors resulted in better discriminant validity than corresponding ICM-CFA factors. Based on cross-lagged autoregressive models of ESEM factors over three occasions, they also found that bully and victim ESEM factors of the APRI were reciprocally related over time—bullies become victims and victims become bullies.

Investigating the psychometric properties of the Physical Self Inventory, Morin and Maïano (2011) similarly showed that ESEM factors were more differentiated than ICM-CFA factors and provided a better fit to the data. This result is important since a repeated observation in physical self concept research (Marsh & Cheng, 2012) is the presence of elevated correlations among global self-worth, physical self-worth, and perceived physical appearance. Morin and Maïano's results showed that these correlations were related to some indicators of perceived physical appearance that also played a key role in defining global and physical self worth. In addition, their results allowed them to pinpoint problems related to negatively worded items that were not apparent with CFA.

Confirmatory vs. Exploratory Factor Analysis

We conceive of ESEM as primarily a confirmatory tool that provides a viable option to the sometimes over-restrictive assumptions of the ICM-CFA model. Ideally, applied researchers should begin with a well-established, a priori factor structure model—consisting of at least the number of factors and the pattern of target and non-target loadings. Tests of such an a priori model are clearly confirmatory in nature. Nevertheless, some post-hoc modifications or exploratory research questions might be required when the model is extended (e.g., partial invariance in tests of measurement invariance or relations with a set of covariates). ESEM studies considered thus far fit into this confirmatory framework and demonstrate that ESEM frequently outperforms corresponding ICM-CFA models even in cases where there is a clear a priori factor model to be tested. However, in applied research settings, researchers sometimes do not have the luxury of a well defined a priori model. One common approach has been to use exploratory EFA to 'discover' an appropriate factor structure and then incorporate this post hoc model into a CFA framework. Clearly this approach blurs the distinction between confirmatory and exploratory factor analysis in a way that may offend purists, but we would not like to automatically reject the appropriateness of

such an approach so long as interpretations are offered with appropriate caution. However, we would like to note that the logical next step following the discovery of the 'best' factor structure based on exploratory EFA might be a confirmatory ESEM rather than CFA as illustrated by two recent studies that we summarise here.

In the first of these studies, Meleddu, Guicciardi, Scalas, and Fadda (2012) explored the factor structure of an Italian version of the Oxford Happiness Inventory. After contrasting factor solutions with one to seven factors based on exploratory EFA, they retained a five-factor solution and showed that it provided a better representation of the data than alternative ICM-CFA models. Then they incorporated this exploratory EFA solution into a confirmatory ESEM framework and showed the five-factor model to be reasonably invariant across gender. Next, they used the ESEM factor correlation matrix to explore the single-factor higher-order structure of the instrument. Similarly, Myers, Chase, Pierce, and Martin (2011) explored the factor structure of the coaching efficacy scale for the head coaches of youth sports teams. After contrasting solutions with one to six factors based on exploratory EFA of ordered categorical indicators, they retained the five-factor ESEM solution, showed its superiority over a comparative ICM-CFA solution, and used ESEM to confirm the invariance of this factor model according to coach gender.

Organization of the examples provided in this chapter

From these illustrations, it is obvious that the main use of ESEM to date has been a psychometric one, as it should be the case since obtaining a well-defined factor structure is a prerequisite to any predictive SEM analyses. This observation can also be explained by the fact that ESEM came to fill an important gap in research, allowing researchers using one of many instruments with a well replicated (i.e., confirmatory) EFA factor structure that did not fit the restrictive ICM-CFA model the possibility to embark on a more rigorous process of scale validation. Indeed, based on a total of eight different long and short instruments

measuring seven different constructs, these illustrations all showed the superiority of ESEM compared to traditional ICM-CFA. However, it is also obvious from these illustrations that the full scope of ESEM possibilities is as large, or even larger since it can be based on EFA factors, CFA factors, or a combination of both, as in SEM or GSEM. It would thus be unrealistic for a single chapter to even attempt such an extensive coverage. Rather, we believe that the core aspect of ESEM lie in the extensive set of psychometric applications (e.g., measurement invariance) that are made available within the ESEM framework. We believe that once the reader masters how to specify, constrain, and more generally model EFA factors within the ESEM framework, most predictive, multilevel, mixture, or other applications of ESEM can easily be deduced from reading the other chapters of this advanced SEM textbook (or any introductory SEM textbook) once the limitations ESEM pose on these are understood. For this reason, we will devote significant space to psychometric applications of ESEM, before moving to a shorter section on predictive applications.

More specifically, the ESEM modeling framework approach will first be presented. Then, psychometric applications of ESEM for the estimation of measurement models, multiple group measurement invariance and longitudinal measurement invariance will be illustrated. Predictive ESEM models will then be briefly presented and these models will build on the preceding psychometric models, with a special attention given to the main limitations of ESEM. Finally, we also present Marsh et al. (2012) EWC generalization of ESEM as a way to circumvent some of the limitations of the ESEM models and provide some illustration of these models. To accompany this chapter, extensive sets of supplemental materials are available online (http://www.statmodel.com/esem.shtml), including all of the input files used to estimate the models described in this chapter, the simulated data file that we used, as well as the input used to generate this data file. It should be noted that, given the scope of this book, we assume readers to be reasonably familiar with EFA, CFA, and SEM,

and to have previously conducted and interpreted such analyses according to current best practices (for a refresher, we suggest Bollen, 1989; Brown, 2006; Byrne, 2011; Cudeck & MacCallum, 2007; Fabrigar, Wegner, MacCallum, & Strahan, 1999; Gorsuch, 1983; Henson & Roberts, 2006; Kahn, 2006; Thompson, 2004; Thompson & Daniel, 1996).

The ESEM Approach

The ESEM model and rotational indeterminacy.

As the objective of the present chapter is not to provide a detailed technical presentation of ESEM, we only summarize selected ESEM features of particular relevance. More details are available in the online supplementary materials (http://www.statmodel.com/esem.shtml). In the ESEM model (Asparouhov & Muthén, 2009; Marsh et al., 2009), there are p dependent variables $\mathbf{Y} = (Y_1, ..., Y_p)$, q independent variables $\mathbf{X} = (X_1, ..., X_q)$, and m latent variables $\boldsymbol{\eta} = (\eta_1, ..., \eta_m)$, forming the following general ESEM model:

$$\mathbf{Y} = \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\eta} + \mathbf{K}\mathbf{X} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{\eta} = \boldsymbol{\alpha} + \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\mathbf{X} + \boldsymbol{\zeta}\,.$$

Standard assumptions of this model are that the $\boldsymbol{\varepsilon}$ and $\boldsymbol{\zeta}$ residuals are normally distributed with mean 0 and variance covariance matrix $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ respectively. The first equation represents the measurement model where $\boldsymbol{\nu}$ is a vector of intercepts, $\boldsymbol{\Lambda}$ is a factor loading matrix, $\boldsymbol{\eta}$ is a vector of continuous latent variables, $\mathbf{K}$ is a matrix of $\mathbf{Y}$ on $\mathbf{X}$ regression coefficients, and $\boldsymbol{\varepsilon}$ is a vector of residuals for $\mathbf{Y}$. The second equation represents the latent variable model where $\boldsymbol{\alpha}$ is a vector of latent intercepts, $\mathbf{B}$ is a matrix of $\boldsymbol{\eta}$ on $\boldsymbol{\eta}$ regression coefficients, $\boldsymbol{\Gamma}$ is a matrix of $\boldsymbol{\eta}$ on $\mathbf{X}$ regression coefficients, and $\boldsymbol{\zeta}$ is a vector of latent variable residuals.

In ESEM, $\boldsymbol{\eta}$ can include multiple sets of ESEM factors defined either as EFA or CFA factors. More precisely, the CFA factors are identified as in traditional SEM where each

factor is associated with a different set of indicators. EFA factors can be divided into blocks of factors so that a series of indicators is used to estimate all EFA factors within a single block, and a different set of indicators is used to estimate another block of EFA factors. However, specific items may be assigned to more than one set of EFA or CFA factors. Assignments of items to CFA and/or EFA factors is usually determined based of a priori theoretical expectations, practical considerations, or, perhaps, post-hoc based on preliminary tests conducted on the data.

In a basic version of the ESEM model including only CFA factors (and thus equivalent to the classical SEM model), all parameters can be estimated with the maximum likelihood (ML) estimator or robust alternatives using conventional identification constraints. However, when EFA factors are posited, a different set of constraints is required to achieve an identified solution (Asparouhov & Muthén, 2009; Marsh et al., 2009). In the first step, an unconstrained factor structure is estimated. Given the need to estimate all loadings, a total of m^2 constraints are required to achieve identification for EFA factors (Jöreskog, 1969). These constraints are generally implemented by specifying the factor variance-covariance matrix as an identity matrix and constraining factor loadings in the right upper corner of the factor loading matrix to be 0 (for the i^{th} factor, $i-1$ factor loadings are restricted to 0). Regarding the ESEM mean structure, the identification is similar to typical CFA: all items intercepts are freely estimated and all latent factor means are constrained to 0 (due to rotational difficulties, the alternative CFA method of constraining one intercept per factor to 0 to freely estimate the latent means is not recommended in ESEM). All of these constraints are built in as the default in the Mplus estimation process; in addition, Mplus uses multiple random starting values in the estimation process to help protect against nonconvergence and local minima. For a detailed presentation of identification and estimation issues, the readers are referred to Asparouhov and Muthén (2009), Marsh et al. (2009, 2010), and Sass and Schmitt (2010).

In the second step, this initial, unrotated solution is rotated using any one of a wide set of orthogonal and oblique rotations (Asparouhov & Muthén, 2009; Sass & Schmitt, 2010). As in any form of EFA, multiple orthogonal and oblique rotation procedures are available in ESEM. A review of these rotational procedures is beyond the scope of the present chapter and extended discussions of alternative rotation procedures are available elsewhere (Asparouhov & Muthén, 2009; Bernaards & Jennrich, 2005; Browne, 2001; Jennrich, 2007; Marsh et al., 2009, 2010; Sass & Schmitt, 2010). In traditional applications of EFA, researchers often choose the rotational criteria on the basis of whether the resulting factors are believed to be orthogonal or correlated (e.g., Fabrigar et al., 1999; Henson & Roberts, 2006). This is clearly unsatisfactory and best practices are evolving (see Sass & Schmitt, 2010), but orthogonal rotations are generally considered to be unrealistic in psychological research. The choice of the most appropriate rotation procedure is to some extent still an open research area in EFA, and even more so in ESEM. However, due to rotational indeterminacy, all forms of rotations have equivalent implications for the covariance structure and thus represent statistically equivalent models.

Here, we focus on Geomin rotation (Browne, 2001; Yates, 1987). It was specifically developed to represent simple structure as conceived by Thurstone (1947), in which cross loadings are ideally small relative to target loadings but not required to be zero as in ICM-CFA . Geomin rotations also incorporate a complexity parameter (ε) consistent with Thurstone's original proposal. As operationalized in Mplus' defaults, this ε parameter takes on a small positive value that increases with the number of factors (Asparouhov & Muthén, 2009; Browne, 2001). Asparouhov and Muthén (2009) recommended estimating ESEM models with varying ε values. Marsh et al. (2009, 2010) generally recommended using an ε value of .5 with complex measurement instruments as the most efficient way of deflating factor correlations, but recently noted that target rotation could be more efficient in some

cases (Marsh, Lüdtke et al., 2011). Until more is known about the conditions of superiority of one method over another, we recommend comparison of results based on alternative rotational procedures or clear arguments for their choice of rotational method.

Although we described the generic ESEM model as starting from an unconstrained factor structure, it is also possible to build in equality constraints in this initial solution that can then be submitted to constrained rotation. The main application for this procedure is for tests of measurement or latent mean invariance across meaningful subgroups of participants or across multiple time-points for the same group of participants in a longitudinal study in order to ensure that the constructs are defined the same way in these different groups, or time points. Thus, in order to test for invariance of sets of ESEM factors, blocks of model parameters can be restricted to be invariant across groups or time points in the estimation of the unconstrained solution and the subsequent rotation. In this way, ESEM easily implements tests of the invariance of factor loadings, item intercepts, uniquenesses, latent variance-covariance and latent means that were not available in conventional implementations of EFA.

Goodness of fit

In applied SEM research, many scholars are seeking universal “golden rules” allowing them to make objective interpretations of their data rather than being forced to defend subjective interpretations (Marsh, Hau, & Wen, 2004). Many fit indices have been proposed (e.g., Marsh, Balla, & McDonald, 1988), but there is even less consensus today than in the past as to what constitutes an acceptable fit. Some still treat the indices and recommended cut-offs as golden rules; others argue that fit indices should be discarded altogether; a few argue that we should rely solely on chi-square goodness-of-fit indices; and many (like us) argue that fit indices should be treated as rough guidelines to be interpreted cautiously in combination with other features of the data. Generally, given the known sensitivity of the chi-square test to sample size, to minor deviations from multivariate normality, and to minor

misspecifications, applied SEM research generally focuses on indices that are sample-size independent (Hu & Bentler, 1999; Marsh, Balla, & Hau, 1996; Marsh et al., 2004; Marsh et al., 2005) such as the Root Mean Square Error of Approximation (RMSEA), the Tucker-Lewis Index (TLI), and the Comparative Fit Index (CFI). The TLI and CFI vary along a 0-to-1 continuum, with values greater than .90 and .95 typically reflecting acceptable and excellent fit to the data, respectively. Values smaller than .08 or .06 for the RMSEA respectively support acceptable and good model fit.

For the comparison of two nested models, the chi-square difference test can be used, but this test suffers from even more problems than the chi-square test for single models that led to the development of other fit indices (see Marsh, Hau, Balla, & Grayson, 1998). Cheung and Rensvold (2002) and Chen (2007) suggested that if the decrease in fit for the more parsimonious model is less than .01 for incremental fit indices like the CFI, then there is reasonable support for the more parsimonious model. Chen (2007) suggested that when the RMSEA increases by less than .015 there is support for the more constrained model. For indices that incorporate a penalty for lack of parsimony, such as the RMSEA and the TLI, it is also possible for a more restrictive model to result in a better fit than a less restrictive model. However, we emphasize that these cut-off values only constitute rough guidelines.

Given the lack of consensus about fit indices, it is not surprising that there is also ambiguity in their application in ESEM and to the new issues that ESEM raises. For example, because the number of factor loadings alone for the EFA factors included in ESEM applications is the product of the number of items and the number of factors, the total number of parameter estimates in ESEM applications can be massively more than in CFA. This feature might make problematic any index that does not control for parsimony (due to capitalization on chance), and yet call into question the appropriateness of controls for parsimony in indices that do. In the meantime, we suggest that applied researchers use a

multifaceted approach based on the integration of a variety of different indices, detailed evaluations of the actual parameter estimates in relation to theory, a priori predictions, common sense, and a comparison of viable alternative models specifically designed to evaluate goodness of fit in relation to key issues.

Data Generation and ESEM Analyses

For present purposes we simulated a multivariate normal data set. This data set includes six items (X_1 through X_6) serving as indicators of two correlated factors, with the first three having their primary loadings on the first factor and the last three having their primary loadings on the second factor, and with two items presenting significant cross loadings on the other factor. These data will be referred to as *Time 1* data. We also simulated a second set of items (Y_1 through Y_6), referred to as *Time 2* data, designed to represent a second measurement point for the *X* items and simulated to have similar properties to the Time 1 data. Two subgroups of participants (n = 3000 and n=1500) were simulated. The simulated data set (ESEM.dat), the inputs codes used to generate the data (data-generation.inp), and all of the (annotated) input codes used in the present chapter are available in online supplemental materials (http://www.statmodel.com/esem.shtml). The population generating model is presented in Figure 1. All analyses were conducted with Mplus 6.1 (Muthén & Muthén, 2010), using the default maximum likelihood (ML) estimator—although other options might be preferable for other situations (e.g., multilevel, ordered-categorical, or missing data) that can be implemented easily in conjunction with ESEM. Following Marsh et al. (2009, 2010) we will rely on an oblique Geomin rotation (the default in Mplus) with an ε value of 0.5.

Insert Figure 1 about here

Illustrating ESEM:

Psychometric Applications Based on the ESEM Measurement Model

Comparison of ESEM and CFA models

The starting point for any ESEM investigation should be to test the a priori factor model and the hypothesis that the ESEM model provides a better fit to the data than a traditional ICM-CFA model[1]. 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 and construct validity of the interpretation of the factor structure. In the current example, CFA and ESEM models were first estimated separately on Time 1 and Time 2 data. The goodness of fit results from these models are reported at the top of Table 1. These results clearly indicate that the ICM-CFA model does not provide an acceptable fit to the data at either time point (significant χ^2; TLI < .95; RMSEA > .08). Conversely, the ESEM model provides an almost perfect fit to the data at both time points (non-significant χ^2; CFI and TLI > .95; RMSEA < .06). It is also instructive to compare parameter estimates based on the ICM-CFA and ESEM solutions, as reported in Table 2. The main difference—in addition to the ESEM cross-loadings—is that the CFA model results in highly inflated factor correlations due to the unrealistic assumption of 0 cross loadings (for Time 1 and Time 2 respectively, r = .728-.709 for CFA vs. .439-.429 for ESEM solution vs. .30 for the population model). Different forms of rotations were compared. These results are reported in the on-line supplementary materials (http://www.statmodel.com/esem.shtml) and support our decision to rely on Geomin rotation with ε = 0.5 with the current data set.

[1] Although it is generally better to guide empirical research from well-supported, a priori hypothesis, in some cases exploratory applications cannot be avoided. An important issue when an ESEM model (then closer to classical EFA applications) is used for purely exploratory purposes is to determine the optimal number of factors required to best represent the data. A brief presentation of criteria that could be used for this purpose is available in the supplemental materials at http://www.statmodel.com/esem.shtml. However, in this chapter we focus on the perhaps more common case where ESEM is used for confirmatory purposes based on a well-defined, a priori theoretical factor structure supported by theory, design, and empirical research.

However, the choice of the optimal rotation criteria remains an open question and different rotational procedures should generally be explored in ESEM studies. For the moment, we recommend that applied researchers should either explore alternative rotation procedures, or rely on a strong rationale for choosing the rotation procedures that they used.

Insert Tables 1 and 2 about here

Measurement invariance: The multiple group approach

Tests of measurement invariance evaluate the extent to which measurement properties generalize over multiple groups, situations, or occasions. Such tests are widely applied in SEM studies (Meredith, 1993; Vandenberg & Lance, 2000) and the evaluation of measurement structures more broadly. Measurement invariance is fundamental to the evaluation of construct validity and generalizability. More specific to our chapter, measurement invariance is an important prerequisite to any form of valid group-based comparison. Indeed, unless the underlying factors are measuring the same construct in the same way and the measurements themselves are operating in the same way (across groups or over time), mean differences and other comparisons might potentially be invalid. Before ESEM, the multigroup tests of invariance were seen as a fundamental advantage of CFA/SEM over EFA. Indeed, traditional EFA approaches were largely limited to descriptive comparisons of the factor loadings estimated separately in each group, while CFA approaches allowed researchers to constrain factor loadings to be the same in multiple groups.

Item intercepts play a key role in tests of measurement invariance, but not in traditional EFA approaches. In particular, an important assumption in the comparison of group means across multiple groups (or over time) is the invariance of a majority of item intercepts, a violation of which is also called *differential item functioning* (DIF; technically, monotonic

DIF). For example, if group differences are not consistent in direction and magnitude across the items associated with a particular latent factor, then the observed differences on the latent factor depend on the mix of items considered. A subset of the items actually used or a new sample of items designed to measure the same factor might give different results.

Following Meredith (1993; also see Millsap, 2011; Vandenberg & Lance, 2000), Marsh et al. (2009, 2010) operationalized a taxonomy of 13 partially nested models varying from the least restrictive model of configural invariance with no invariance constraints to a model of complete invariance positing strict invariance of the factors loadings, items intercepts, and items uniquenesses as well as the invariance of the latent means and of the factor variance-covariance matrix (see Table 3). We illustrated this multigroup ESEM approach to invariance with our simulated data. As the results are highly similar across time points (and were specified to be similar in the population generating model), multiple group tests of invariance were only conducted on Time 1 data (i.e., variables X_1 through X_6) due to space limitations (readers are invited to conduct parallel tests with Time 2 data; files for doing so are contained in the supplemental materials at http://www.statmodel.com/esem.shtml). It should also be noted that when the multiple groups reflect a factorial combination of grouping variables (e.g., 3 age levels × 2 genders = 6 groups), it is possible to discern non-invariance due to the interaction of the two grouping variables as well as the "main" effects of each grouping variable considered separately. The results from the 13 steps of multiple group measurement invariance tests are reported in Table 1, shown previously. To conserve space, we will focus our presentation of results to the classical models of configural (Model 1), weak (Model 2), strong (Model 5), strict (Model 7) invariance as defined by Meredith's (1993) seminal paper, before exploring the invariance of the latent variance-covariance matrix and mean structure.

Insert Table 3 about here

Configural invariance (Model 1). This taxonomy begins with a model having no invariance of any parameters across the groups, or *configural invariance* (Model 1). Like the traditional EFA approach, factor structures are freely estimated separately in each group, with only the number of factors being the same in all groups. This model is usually identified by using the latent standardization method of fixing the latent variances to 1 and the latent means to 0 in all groups so as to freely estimate all factors loadings and items intercepts. Model 1 provides a baseline of comparison for the remaining models in the taxonomy that are nested under it (i.e., parameters in the subsequent, more restricted model are a subset of parameters of Model 1). This baseline model with no invariance constraints provides a satisfactory level of fit to the data (non-significant χ^2; CFI and TLI > .95; RMSEA < .06).

Weak measurement invariance (Model 2). Weak measurement invariance tests whether the factor loadings are the same in both groups. This is an important test, as all subsequent models in the taxonomy assume the invariance of factor loadings. When the loadings are fixed to equality across groups, the variances in all groups except in an arbitrarily selected reference group can be freely estimated. The test of weak measurement invariance is based on the comparison of Model 2 with Model 1. For our simulated data, this comparison results in a non-significant chi-square difference test, equivalent CFIs and TLIs (to two decimals), and a slight improvement in the RMSEA (that adjusts for model parsimony). These results provide strong support for weak measurement invariance. It should be noted that tests of weak measurement invariance are critical in ESEM as it is currently not straightforward to specify a model with partial invariance of factor loadings (but see the ESEM-within-CFA method, described later).

Strong measurement invariance (Model 5). Model 5 tests the strong measurement invariance of the model and requires that the indicator intercepts—in addition to the factor

loadings—are invariant over groups. This is an important test since it justifies the comparison of latent means. When the intercepts are constrained to equality, the latent means can be freely estimated. The typical approach is to constrain the means in a selected referent group to be zero, and to use this as a basis of comparison for the remaining means that are freely estimated. Although the crucial comparison for strong measurement invariance is between Models 2 and 5, the invariance of items' intercepts can also be tested by the comparison of any pair of models differing only in regard to intercept invariance (Model 2 vs. Model 5; Model 3 vs. Model 7; Model 4 vs. Model 8; Model 6 vs. Model 9) and convergence or divergence of conclusions can be inspected for further information. For our simulated data, all of these comparisons resulted in significant chi-square difference tests, and in substantial decreases in CFI and TLI and an increase in RMSEA that exceed the recommended cut-offs. This shows that the hypothesis of strong measurement invariance does not hold for all items.

When tests of full invariance fail, it is typical to consider tests of partial invariance in which the invariance constraints for one or a small number of parameters are relaxed. For our simulated data, examination of the modification indices associated with these models suggests that the lack of invariance is limited to item 3 (consistent the population generating model). A series of partial invariance models were thus considered (those labeled "p" in Table 1) in which the invariance of the item 3 intercept was relaxed while invariance constraints were retained for the remaining intercepts. When the previously noted pairs of models were compared but relying on partial, rather than complete, invariance of the items intercepts, the results support that the remaining items are invariant across groups (non-significant $\Delta\chi^2$ and ΔTLI, ΔCFI, and ΔRMSEA < .01). When there are more than two groups, partial invariance could be limited to a subset of groups.

For instrument construction, DIF might be a sufficient basis for dropping an item (if there is a large pool of items with invariant factor loadings and intercepts from which to

choose in the pilot version of the instrument). However, the post-hoc construction of models with partial invariance should generally be viewed with caution. When the sample size is small or non-representative, there is the danger of capitalizing on idiosyncrasies of the sample. When the number of target items measuring each factor is small, there is the danger that the invariance found for a subset of the items is not generalizable to the population of all possible items that could have been picked to assess this specific construct. Here, for example, the observed latent mean differences between groups are a function of the items that do have invariant intercepts and are assumed to generalize to other sets of appropriate items.

Strict measurement invariance (Model 7). Model 7 requires invariance of items' uniquenesses in addition to the invariance of factor loadings and intercepts. The invariance of uniquenesses is a prerequisite to the comparison of manifest scores based on factor scores but is also relevant and interesting in its own right as a test of the generalizability of measurement error across groups. Still, tests of uniqueness are not as critical as other tests, as long as one relies on latent variable methodologies. Thus, if there is clear support for the invariance of uniquenesses, it is fine to impose this constraint; however, if the invariance of uniquenesses is not supported, comparisons of latent means are still justified without this constraint. Although the crucial comparison for strict measurement invariance is between Model 5 and Model 7 (or 5p and 7p in the case of partial invariance), the invariance of items' uniquenesses can also be tested by the comparison of any pair of models differing only in regard to uniquenesses' invariance (Model 2 vs. Model 3; Model 4 vs. Model 6; Model 5 vs. Model 7; Model 8 vs. Model 9; Model 10 vs. Model 11; Model 12 vs. Model 13). Comparisons of each of these pairs of models result in a non-significant chi-square difference test and in equivalent CFI, TLI, and RMSEA values (to two decimals). Hence, there is support for partial strict measurement invariance – the partial label is determined by the non-invariant intercept.

Factor variance-covariance invariance (Model 4). Model 4 tests the invariance of the

factor variance-covariance matrix—in addition to factor loadings. Although the invariance of each factor variance and covariance term can be tested separately in CFA, this is not routinely possible with ESEM (due to complications in rotation; but see the ESEM-within-CFA method described later). Tests of the invariance of the factor variance-covariance matrix (Model 4) are highly relevant to many substantively important issues, even more so than the comparison of latent means in some applications. Nevertheless, for purposes of testing group means, the invariance of the variance-covariance matrix is not necessary and does not represent an essential step in tests of measurement invariance. For our simulated data, the critical comparison is between Model 2 and Model 4 (or other pairs of models differing in respect to the invariance of the variance-covariance matrix). Although some of these comparisons result in statistically significant chi-square differences, none of them result in changes in CFI, TLI, or RMSEA that exceeds the recommended thresholds. These results thus provide reasonable support for the invariance of the factor variance-covariance matrix.

Invariance of latent means (Models 10 through 13). The final four models (Model 10 through Model 13) in the taxonomy all constrain mean differences between groups to be zero – in combination with the invariance of other parameters. In order for these tests to be interpretable, it is essential that there is support for the invariance of factor loadings and item intercepts (or at least partial invariance for a majority of items per factor, for further discussion see Byrne, Shavelson, & Muthén, 1989), but not item uniquenesses or the factor variance-covariance matrix. The most appropriate model will depend on the results of earlier invariance tests. In general, it is advisable to test the most constrained model that is justified by earlier tests. Hence, if there is support for the invariance of item uniquenesses and the factor variance-covariance matrix (as well as factor loadings and item intercepts) then Model 13 should be compared to Model 9. However, if there were invariance support for neither item uniquenesses nor the factor variance-covariance matrix (but there was support for factor

loading and item uniquenesses), then the appropriate test would be the comparison of Model 10 and Model 5. For our simulated data, tests of Model 10 through Model 13 all lead to the conclusion that latent means differ systematically across groups (significant $\Delta\chi^2$ and ΔTLI, ΔCFI, and ΔRMSEA > .01). These comparisons show that, when group 1 latent means are constrained to be zero, group 2 latent means vary between .513 and .633 on factor 1 and .476 and .517 on factor 2 – all close to the true population values of .50.

Summary. The multiple group approach provides a very general and elegant framework for tests of measurement invariance and latent mean differences when the grouping variable has a small number of discrete categories and the sample size for each group is reasonable. The extension of ESEM to incorporate this multiple group approach is one of the most important applications of ESEM. Nevertheless, this multiple group approach might not be practical for variables that are continuous (e.g., age), for studies that evaluate simultaneously many different contrast variables (e.g., age, gender, experimental/control) and their interactions, or when sample sizes are small. In such situations, a more parsimonious MIMIC approach (to be presented later) might be appropriate.

Measurement invariance: The longitudinal approach

Essentially the same logic and the same taxonomy of models can be used to test the invariance of parameters across multiple occasions for a single group. One distinctive feature of longitudinal analyses is that they should normally include correlated uniquenesses (CUs) between responses to the same item on different occasions (see Jöreskog & Sörbom, 1977; Marsh, 2007; Marsh & Hau, 1996). When the same items are used on multiple occasions, the uniqueness component associated with each item from one occasion is typically positively correlated with the uniqueness component associated with the same item on another occasion. Failure to include these CUs generally results in biased parameter estimates. In particular, test-retest correlations among matching latent factors are systematically inflated, which can

then systematically bias other parameter estimates and may even result in improper solutions such as a non-positive definite factor variance-covariance matrix or estimated test-retest correlations that exceed 1.0 (e.g., Marsh, Martin, & Hau, 2006; Marsh, Martin, & Debus, 2001). Interestingly, the inclusion of CUs is another option within ESEM that was not typically possible in traditional EFAs.

For our simulated data, preliminary analyses were conducted to show that models including CUs (i.e., X_1 with Y_1, X_2 with Y_2, etc) provided a better fit to the data for both CFA and ESEM models (see first four models in Table 4). Interestingly, comparison of these four models demonstrated that the CFA model with CUs was acceptable according to some criteria (CFI, TLI > .95; RMSEA =.058)—even though all indices of fit were substantially better for the ESEM model (CFI, TLI > .99; RMSEA =.014). However, inspection of the factor correlations based on these four models, which are reported in Table 5, reveals that CFA factor correlations were systematically higher (and inflated relative to the known population parameters for our simulated data) when compared to the ESEM factor correlations. Of particular relevance, the test-retest correlations are systematically higher for models that do not contain CUs (and inflated relative to the known factor correlations for our simulated data) than for models that do. On this basis, we evaluated longitudinal invariance in relation to ESEM models that included CUs.

Insert Tables 4 and 5 about here

The results from the 13 steps of longitudinal measurement invariance tests are reported in Table 4. As the results from the longitudinal invariance tests closely parallel those from the multiple group invariance tests, they will not be described in detail. These results support the complete longitudinal invariance of the factor loadings, of the factor variance-covariance

matrix, and of the items' uniquenesses. However, the results did not support the complete invariance of the items' intercepts. Examination of the modification indices associated with the various models allowing for the verification of items' intercept invariance suggests that this lack of invariance is limited to item 6 (consistent with the population generating model). A series of partial invariance models were thus estimated (labeled "p" in Table 4) in which the invariance of item 6 intercept was relaxed while the remaining intercepts were constrained to be invariant. Lastly, the final four models from the taxonomy all converged in rejecting the longitudinal invariance of the factor means. These models revealed that, when Time 1 latent means are constrained to be zero, Time 2 latent means vary between .465 and .508 on factor 1 and on .448 and .506 on factor 2, close to the population values of .50.

Summary. The longitudinal approach to invariance closely parallels the multiple group approach and provides a very flexible framework for the examination of the invariance assumptions inherent in longitudinal analyses. Due to space limitations, we do not pursue tests of multiple group longitudinal invariance, which would be the next logical step. Conducting these tests would involve multiplying by three the sequence of 13 steps proposed in Table 3 as these models allow for the testing of the invariance of the models parameters across groups, time period, and groups × time periods. In addition, new models could be added to this taxonomy (e.g., Marsh et al., 2010) to investigate the invariance of the longitudinal CUs. However, these tests can be conducted with our simulated data and we suggest that interested readers explore these possibilities.

Illustrating ESEM:

Predictive Applications Based on the ESEM Structural Model

The MIMIC approach.

The MIMIC model (Jöreskog & Goldberger, 1975; Marsh, Ellis, Parada, Richards, &

Heubeck, 2005; Marsh, Tracey, & Craven, 2006; Muthén, 1989) is a multivariate regression model in which latent variables are regressed on observed predictors (see Figure 2). In addition, when the latent variables have multiple indicators, the MIMIC model can also be extended to test potential non-invariance of item intercepts, that is, DIF (technically, monotonic DIF). In that specific case, the MIMIC model has important advantages over the multiple group approach, but also some limitations. Particularly in applied research based on often modest sample sizes, the MIMIC model is much more parsimonious and does not require the separate estimation of the model in each group. Also, it allows researchers to consider multiple independent variables that would typically become unmanageable in multiple group analyses. A particularly important feature of the MIMIC model is that it allows researchers to consider continuous predictors (e.g., age, income, pretest scores) that cannot be evaluated in the multiple group approach without recoding them to form a small number of discrete groups. It is also possible to consider a combination of continuous and categorical predictors (and their interaction) in the same MIMIC model. However, while the MIMIC model is able to test DIF, it does not test the invariance of factor loadings (non-monotonic DIF), or the factor variance-covariance matrix (and implicitly assumes their invariance). We focus here on the use of the MIMIC model as a way to investigate monotonic DIF as it provides a more complete illustration of the full flexibility of the MIMIC approach, but note that this section more generally provides a generic example of how to relate ESEM factors to any number of observed variables (and their interactions).

Insert Figure 2 about here

In the MIMIC approach, monotonic DIF can be evaluated by the comparison of three models. The first (null effect) MIMIC model posits that the predictor variables have no effect

on the latent variables and items intercepts (i.e., paths from predictors to latent factors and their indicators—the full lines and the dotted-dashed lines in Figure 2—are constrained to be zero). The second (saturated) MIMIC model has paths from each predictor variable to all item intercepts (i.e., the dotted-dashed lines in Figure 2), but not the latent factors. The third (invariant intercept) MIMIC model has freely estimated paths from the predictor variables to the latent factors (i.e., the full lines in Figure 2), but paths to item intercepts are all constrained to be zero. The comparison of Model 1 with Models 2 and 3 tests whether there are any effects of the predictors, the comparison of Model 1 and Model 3 tests whether the predictors have an effect on the latent variables, while the comparison of Model 2 and Model 3 tests whether the effects of the predictor variables on individual items can be fully explained in terms of effects on the latent factors. If Model 2 fits substantially better than Model 3, then there is evidence of monotonic DIF (i.e., non-invariance of intercepts). In this case, it might be appropriate to pursue partially invariant models in which the invariance constraint is relaxed for some item intercepts (but see previous discussion).

For our simulated data (see MIMIC Models in Table 1 and on-line materials), we again focus on the Time 1 results in order to conserve space (but note that Time 2 results are very similar and can easily be tested with the data set provided online). The MIMIC null effect model, in which the grouping variable is posited to be unrelated to the ESEM factors or the items, failed to provide an acceptable fit to the data (significant χ^2; TLI < .95; RMSEA > .08). This suggests that at least some effects of the predictor variable should be expected. Indeed, the saturated MIMIC model did provide a satisfactory fit to the data (non-significant χ^2; CFI and TLI > .95; RMSEA < .06) and a substantial improvement over the null effect model. The third (intercept invariant) MIMIC model (i.e., in which the grouping variable is only allowed to predict the latent factor scores but not the items) failed to provide an acceptable fit (significant χ^2; TLI < .95; RMSEA > .08), suggesting DIF. Examination of the modification

indices associated with this model indicates that DIF was mainly associated with item 3. Allowing for direct effects of the predictor on item 3, in addition to its effects on the ESEM factors, results in a satisfactory fit to the data (non-significant χ^2; CFI and TLI > .95; RMSEA < .06) and in a fit that is comparable to the fit of the saturated MIMIC model [$\Delta\chi^2$ (df) = 4.607 (3), $p > .05$; ΔTLI and ΔCFI < .01]. Detailed results from this model reveal that participants' levels on factor 1 ($\hat{\beta}$ = .523, $p < .001$), factor 2 ($\hat{\beta}$ = .489, $p < .001$), and item 3 ($\hat{\beta}$ = .369, $p < .001$), tend to be higher in the second group.

Summary. The MIMIC approach to measurement invariance provides a generic framework for testing the relations between any number of observed continuous or categorical predictors, and their interactions, as well as a very powerful alternative to the multiple group approach for tests of DIF when the sample size is small, when the predictors are continuous or include many categories, when there are multiple predictors, and when interactions among predictors are considered. However, the MIMIC approach is limited in that it assumes the invariance of factor loadings and uniquenesses, but it does not allow for the verification of these assumptions. To further increase the flexibility of the MIMIC approach, Marsh et al. (2006) proposed a hybrid approach in which multiple group models and the MIMIC approach are combined for greater precision in the investigation of measurement invariance issues. So long as the two approaches converge to similar interpretations, there is support for the construct validity of these interpretations. More recently, Marsh, Nagengast, and Morin (2012) extended this approach to ESEM, including tests specifically designed to investigate the loss of information due to categorizing continuous variables where a MIMIC model is separately estimated in each of the separate groups.

Autoregressive cross lagged ESEM models

An obvious extension of the MIMIC approach that we just presented would be to use

latent variables to predict other latent variables. Rather than to present this simpler scenario, that can easily be pursued with the data set that is provided online, we present one final model to better highlight the flexibility of the ESEM approach in a way that might not be obvious to the reader. An important question that can be pursued in longitudinal research is related to the direction of the relations between two constructs over time (e.g., Marsh & Yeung, 1998; Morin, Maïano, Nagengast, Marsh, Morizot, & Janosz, 2011). These questions can be investigated in the context of autoregressive cross lagged models (Jöreskog & Sörbom, 1977; Marsh & Grayson, 1994), where each variable is expressed as an additive function of the preceding values on both variables (here Factors 1 and 2) and a random error. See Figure 3 for details of this model, as well as the syntax used to estimate this model (ARCLM-from7p.inp) in the on-line materials. This model was built from the longitudinal Model 7p (i.e., partial strict measurement invariance) since measurement invariance of the constructs over time is a prerequisite to longitudinal analyses in order to ensure that the repeated measures estimate the same constructs over time. The results from this model reveal strong "horizontal" (i.e., test-retest) effects where each variable mostly predicted itself over time ($\hat{\beta}_{F1_t,F1_{t-1}}$ = .527; $p < .001$; $\hat{\beta}_{F2_t,F2_{t-1}}$ = .425, $p < .001$). Small longitudinal effects going from Factor 2 to Factor 1 were also apparent ($\hat{\beta}_{F1_t,F2_{t-1}}$ = -.069, $p \leq .001$), but the reciprocal effects of Factor 1 to Factor 2 were nonsignificant ($\hat{\beta}_{F2_t,F1_{t-1}}$ = .027, $p > .05$), suggesting that the direction of influence goes from Factor 2 to Factor 1 and is negative once the stability of each process is controlled for in the model.

Insert Figure 3 about here

This model can easily be extended to incorporate additional measurement points (see Marsh, Nagengast, et al., 2011) and can also serve as a starting point for the estimation of any

predictive relationships between latent variables in the ESEM framework. It can also be combined with the MIMIC model so as to estimate the relations between observed and latent variables. For instance, we re-estimated this model while controlling for the effects of the grouping variable on the various latent factor, which is akin to estimating the previous autoregressive cross lagged model with the grouping variable used as a MIMIC-like predictor of the latent factors (see ARCLM-cont-from7p.inp). This model thus provides a straightforward extension of the previous one when applied researchers want to estimate predictive relationships above and beyond the effects of some confounding variables (e.g., gender) used as control. Not surprisingly given our population model, the overall pattern of results obtained from this example is not changed by the addition of this additional control ($\hat{\beta}_{F1_t,F1_{t-1}}$ = .494; $p < .001$; $\hat{\beta}_{F2_t,F2_{t-1}}$ = .400, $p < .001$; $\hat{\beta}_{F1_t,F2_{t-1}}$ = -.090, $p \leq .001$; $\hat{\beta}_{F2_t,F1_{t-1}}$ = -.008, $p > .05$) and show significant effects of the grouping variables on all latent factors ($\hat{\beta}_{F1_{t1},G}$ = .287, $p < .001$; $\hat{\beta}_{F1_{t2},G}$ = .150, $p < .001$; $\hat{\beta}_{F2_{t2},G}$ = .240, $p < .001$; $\hat{\beta}_{F2_{t2},G}$ = .166, $p < .001$).

Extending ESEM:

The ESEM-Within-CFA (EWC) Approach and Illustrations

The ESEM approach is very flexible, but its current operationalization still presents some limitations when compared to CFA and SEM models. For instance, with standard ESEM models it is impossible, or at least very difficult, to: (a) fit a higher order factor on a set of first order ESEM factors, which also means that fully latent curve models cannot be estimated from longitudinal sets of ESEM factors, (b) evaluate mixture or factor mixture models, and (c) constrain ESEM latent means in multiple-group models, for example, 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 and gender interactions (see Marsh,

Nagengast, & Morin, 2012). What is perhaps the most worrisome current limitation of ESEM, however, is that all of the factors forming a set of ESEM factors need to be simultaneously related or unrelated to other variables in the model. For instance, if we take the MIMIC model used previously, one cannot use the grouping variable to predict a single ESEM factor but not the other, or use a single ESEM factor to predict an outcome. Similarly, the invariance of the factor variance-covariance matrix can only be tested in an all-or-none fashion. Applied researchers are not currently able to test the invariance of selected factor variances and covariances separately, or the relations between selected ESEM latent factors and other variables. Marsh, Nagengast, and Morin (2012) proposed a generalization of ESEM they called ESEM-Within-CFA (EWC) as a way to circumvent some of these problems.

EWC is based on an extension of an initial proposal by Jöreskog (1969; also see Muthén & Muthén, 2009, slides 133-146) that was designed to provide standard errors for EFA parameter estimates and greater flexibility in specifying factor structures—important limitations of EFA at that time. However, our extension of this earlier approach provides a possible solution to some of the aforementioned limitations of ESEM. The EWC model must contain the same number of restrictions as the ESEM model, that is, m^2 restrictions where m=number of factors (see earlier discussion). To achieve these restrictions, Jöreskog (1969) and Muthén and Muthén (2009) initially proposed that researchers select, based on preliminary EFAs, indicators with near-zero loadings on all but the latent constructs that they are designed to measure. 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 ESEM model. In our EWC approach (Marsh et al., 2011) we offered an amendment to this approach for greater precision. The EWC model is estimated according to the following steps:

(1) In preliminary analyses, compare ESEM and CFA models with regard to goodness of fit and parameter estimates. If the ESEM solution is not clearly superior, ESEM or EWC models should not be pursued and more parsimonious CFA models should be retained.

(2) If preliminary ESEM models are better than CFA models, a complete ESEM analysis should be done to identify the best model with regard to goodness of fit and substantive interpretations.

(3) If there is a need to conduct an additional analysis that cannot be easily implemented within the ESEM framework but can be estimated with CFA models, then all parameter estimates from the final ESEM solution should be used as starting values to estimate the EWC model.

(4) Since a total of m^2 constraints need to be added for the EWC model to be identified, selected parameter estimates are fixed to the values obtained from the ESEM solution:

(i) The m factor variances: by default these parameters are fixed to be 1 in a single-group ESEM solution or for the first group of a multiple group solution.

(ii) A referent indicator is selected for each factor that has a large (target) loading for the factor it is designed to measure and small (non-target) cross-loadings. Then, for purposes of identification, these small cross loadings are fixed to their estimated values from the ESEM solution (i.e., do not allow these values to be freely estimated). Unlike the classical method of fixing these cross-loadings to zero, this approach routinely provides an exact match to the ESEM solutions in terms of parameter estimates, and highly similar standard errors estimates (that might however be slightly inflated).

(iii) For all other parameter estimates, the pattern of fixed and free estimates should be the same as in the selected ESEM solution. Thus, if the parameter is free in the ESEM solution it should be free in EWC and if the parameter is fixed or constrained in ESEM it should also be fixed or constrained in the same way in the EWC solution.

(iv) It should be noted that the mean structure from the EWC 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 EWC solution will have the same degrees of freedom and, within rounding error, the same chi-square, goodness of fit statistics, and most importantly parameter estimates as the ESEM solution. Standard errors will also be highly similar, but might be slightly inflated, suggesting that caution still needs to be exerted in the interpretation of marginally non-significant results. In this sense, it is equivalent to the ESEM solution. Importantly, the researcher has more flexibility in terms of how to constrain or further modify the EWC model (as it is a true CFA model) than with the ESEM model upon which it is based. Interestingly, the EWC approach could be pursued in the present investigation. The fit statistics from the EWC model as estimated on Time 1 data are identical to the results from the ESEM model [χ^2 (df) = 3.085 (4), $p > .05$; CFI = 1.000; TLI = 1.000; RMSEA = .000]. The input (see EWCTime1.inp) is available in the on-line materials.

The EWC model is a convenient way of implementing a specific rotated ESEM solution within a conventional CFA model that allows more flexibility for further analysis than the original ESEM model. However, the initial ESEM model is needed to specify the EWC model. By allowing the researcher to import a well-tested and rotated ESEM measurement model to an even more flexible CFA framework for further analysis, EWC represents a useful complement to ESEM. For example, the EWC approach can be used to extend ESEM to applications that could not easily be tested with the traditional ESEM approach, such as conducting:

- Tests of the partial invariance of factor loadings (see EWC_partload.inp in the on-line materials for an illustration of how to specify this model);

- Tests of invariance separately for the factor variances and covariance (see EWC_var inv.inp in the on-line materials for an illustration);
- Tests of time × grouping variable latent mean differences, implementing contrasts to define main and interaction effects (see EWC_MG_L_7p.inp in the on-line materials). For an illustration of this method applied to more than two measurement points so as to represent nonlinear trajectories, we refer the interested reader to illustrations and syntax provided by Marsh, Nagengast, et al. (2011).

The above are relatively straightforward applications of existing SEM approaches that are easily implemented into the EWC framework so we decided not to further elaborate on them to conserve space. The interested reader can consult the online supplements for the annotated inputs files that may be used to estimate these models. Rather, we focus on the perhaps more common case of longitudinal mediation models, a final predictive example to illustrate how the EWC approach can help to circumvent some limitations of ESEM.

Mediation occurs when some of the effects of an independent variable (IV) on the dependent variable (DV) can be explained in terms of another mediating variable (MV) (Marsh, Hau, Wen, Nagengast, & Morin, in press). A mediator is thus an intervening variable accounting for at least part of the relation between a predictor and an outcome such that the predictor influences an outcome indirectly through the mediator. As such, the temporal ordering of the sequence IV→MV→DV is particularly important in mediation testing. Tests of longitudinal mediation based on latent variables have recently received increased scientific attention (Cole & Maxwell, 2003; MacKinnon, 2008; Selig & Preacher, 2009) as a method of choice to a fuller understanding of developmental processes. These tests are complex and are not all easily implemented in the regular ESEM framework where all factors from a set of ESEM factors need to be simultaneously related to other variables. For instance, consider the case where four predictors (defined as a set of EFA factors) are used to predict an outcome

(defined as a single factor) indirectly via the action of two mediators (defined as a set of EFA factors) so that the first two predictors exert their effect via the first mediator and the last two predictors exert their effects via the second mediator. Mediation models taking this form are common in theoretically-grounded applied research (e.g., Loose, Régner, Morin, & Dumas, in press). However, these models cannot be estimated within the ESEM framework where all predictors need to be related to all mediators, which in turn need to be all related to the outcomes. EWC provides an interesting alternative to ESEM in such cases.

However, this is not the main limitation of ESEM for tests of mediation. Indeed, it is now well documented that bootstrapped confidence intervals represent the most efficient manner of testing the significance of indirect effects (represented as the product of the IV→MV and the MV→DV path coefficients) (e.g., Cheung & Lau, 2008; MacKinnon, Lockwood, & Williams, 2004). Unfortunately, bootstrapping still cannot be implemented in ESEM, but can easily be in EWC.

Moreover, the present data set was simulated with only two measurement points, creating another challenge when one wants to estimate longitudinal mediation models. In fact, with only two measurement points, mediation models fully ordered in time are not possible (unless, naturally, there is an inherent ordering of the variables—such as stable background variables like gender and ethnicity). The closest approximation would be to estimate difference scores between the two measurement points of one or both variables (MacKinnon, 2008; Selig & Preacher, 2009). For illustration purposes, suppose that we hypothesized that the relations between Time 1-Factor 2 (T1F2) and Time 2-Factor 2 (T2F2) would be mediated by changes in Factor 1 (CF1) occurring between Time 1 and Time 2. This model is illustrated in Figure 4. Note that this specific ordering of the effects is consistent with the results from the autoregressive cross-lagged models estimated in the previous section. Difference scores reflecting change occurring over time in specific latent variables

can easily be implemented in a latent variable framework as latent difference scores, but represent a higher order factor that cannot be implemented with ESEM. So we estimated this model using EWC (see long-med-change-from7p.inp and long-med-change-from7p-boot.inp in the on-line supplemental materials) starting from longitudinal Model 7p. The obtained results show significant partial mediation in which T1F2 significantly predicts T2F2 ($\hat{\beta}_{T2F2,T1F2}$ = .540; p < .001) and CF1 ($\hat{\beta}_{CF1,T1F2}$ = -.084; p < .001), which in turns also predicts T2F2 ($\hat{\beta}_{T2F2,CF1}$ = .326; p < .001). Not surprisingly, the indirect effect of T1F2 on T2F2 as mediated by CF1 is negative and significant as indicated from its bootstrapped 95% confidence interval that excludes 0 (indirect effect = -.028, 95%CI = -.048/-.013), showing that higher levels of T1F2 predict lower changes in F1 between time 1 and time 2 (CF1), while these changes predict higher levels in T1F2.

Insert Figure 4 about here

Conclusion

For pedagogical purposes, this chapter relied on a relatively simple simulated data set. This data set can be freely used for the initial stages of ESEM teaching and training, as the decisions to be made remain relatively simple and facilitated by knowing the real population parameter values. However, real data sets are often messier and involve more complicated decisions. As a next step, we recommend that interested readers consult previous applications of ESEM to real data sets that have been described at the start of this chapter to see how the range of possibility illustrated in this chapter can be implemented with real data sets in order to answer substantively important research questions.

Clearly, an important advantage of ESEM is to allow for a more appropriate

representation of factor correlations due to the non-imposition of arbitrary zero cross-loadings—or at least a more systematic approach to the inclusion of cross-loadings into a measurement model than traditional CFA. ICM-CFA apparently systematically inflates the size of correlations among the latent factors when cross-loadings are present (Marsh, Lüdtke, et al., 2011). Indeed, the only way that these cross-loadings can be represented is by inflating the size of correlations. In relation to psychological and social science research more generally this can represent a particularly serious problem because it undermines support for: (a) the multidimensional perspective that is the overarching rationale for many psychometric instruments, (b) the discriminant validity of the factors that form these instruments, (c) the predictive validity of the factors due to multicollinearity, and (d) the usefulness of the ratings in providing diagnostic feedback to the persons being evaluated. We suggest that similar phenomena are likely to occur in most applications where ICM-CFA models are inappropriate. Conversely, allowing for cross-loadings when none are required, although it may result in the over parameterization of the model, is unlikely to result in a negative bias in factor correlations.

In this chapter, we compared factorial solutions generated from ESEM relative to ICM-CFA. Guided by the taxonomy of invariance models, we also demonstrated the utility of ESEM for tests of multiple-group and longitudinal tests of measurement invariance based on the Marsh et al. (2009, 2010) 13-model taxonomy, which can also be applied in traditional ICM-CFA studies. However, to the extent that the ESEM solution provides an acceptable fit to the data and the CFA solution does not, then the appropriateness of the taxonomy for CFA models is dubious. In this respect we present the ESEM as a viable alternative to CFA, but do not argue that it should replace CFA. Still, ESEM should generally be preferred to ICM-CFA when the factors are appropriately identified by ESEM, the goodness of fit is meaningfully better than for ICM-CFA, and factor correlations are meaningfully smaller than for ICM-

CFA. Furthermore, based on Marsh's (2007; Marsh et al., 2005) suggestion that almost no multidimensional psychological instruments widely used in practice provide an acceptable fit in relation to an a priori ICM-CFA structure, we suspect that ESEM is likely to generate better factorial solutions. In this situation, we suggest that advanced statistical strategies such as multi-group tests of measurement invariance, MIMIC models, and even latent growth models in many applications are more appropriately conducted with an ESEM approach than with a traditional ICM-CFA approach. To illustrate this point, we also showed that autoregressive cross lagged models and change-score based longitudinal mediation models could be estimated in ESEM, or within the complementary EWC method. While we believe that the results of existing research provide considerable support to ESEM, there have been only few large-scale applications of ESEM in applied research settings. Clearly there is need for further research based on the application of ESEM to other areas in psychology, education, and the social sciences more generally. Based on results from existing ESEM research we recommend that the psychometric evaluation of psychological assessment instruments should routinely apply ESEM and juxtapose the results with corresponding CFA models that are traditionally used.

References

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

Bagozzi, R. P., Fornell, C., & Larcker, D. (1981). Canonical correlation analysis as a special case of a structural relations model. *Multivariate Behavioral Research, 16,* 437-454.

Bernaards, C. A., & Jennrich, R. I. (2005). Gradient projection algorithms and software for arbitrary rotation criteria in factor analysis. *Educational and Psychological Measurement, 65,* 676-696.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

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

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

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

Browne, M.W ., Cudeck, R., Tateneni, K., & Mels, G. (2010*). CEFA: Comprehensive Exploratory Factor Analysis, version 3.04*. Available at: http://faculty.psy.ohio-state.edu/browne/software.php

Byrne, B. M. (2011). *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. Mahwah, NJ: Routledge.

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.

Carroll, J. B. (1953). An analytical solution for approximating simple structure in factor analysis. *Psychometrika, 18,* 23-38.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*, 464-504.

Cheung, G. W., & Lau, R. S. (2008). Testing mediation and suppression effects of latent variables: Bootstrapping with structural equation models. *Organizational Research Methods, 11*, 296-325.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 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.

Cohen, J. (1968). Multiple regression as a general data-analytic system. *Psychological Bulletin, 70*, 426-443.

Cole, D. A., & Maxwell, S. E. (2003). Testing meditational models with longitudinal data: Questions and tips in the use of structural equation modeling. *Journal of Abnormal Psychology*, *112*, 558-577.

Cudeck, R., & MacCallum, R. C. (Eds.). (2007). *Factor analysis at 100: Historical developments and future directions*. Mahwah, NJ: Erlbaum.

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

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*, 272-299.

Fan, X. (1997). Canonical correlation analysis and structural equation modeling: What do

they have in common? *Structural Equation Modeling: A Multidisciplinary Model*, *4*, 65-79.

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.

Graham, J. M. (2008). The general linear model as structural equation modeling. *Journal of Educational and Behavioral Statistics*, *33*, 485-506.

Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educational and Psychological Measurement, 66*, 393-416.

Hessen, D. J., Dolan, C. V, & Wicherts, J. M. (2006). Multi-group exploratory factor analysis and the power to detect uniform bias. *Applied Psychological Research, 30*, 233-246.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*, 1-55 .

Jennrich, R. I. (2007). Rotation methods, algorithms, and standard errors. In R. Cudeck & R. C. MacCallum (Eds.) *Factor Analysis at 100: Historical Developments and Future Directions* (pp. 315-335). Mahwah, NJ: Erlbaum.

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

Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, *10*, 631-639.

Jöreskog, K. G., & Sörbom, D. (1977). Statistical models and methods for the analysis of longitudinal data. In D. J. Aigner, & A. S. Goldberger (Eds.), *Latent variables in socio-economic models* (pp. 285-325). Amsterdam, NL: North-Holland Publishing.

Kahn, J. H. (2006). Factor analysis in counseling psychology research, training, and practice:

Principles, advances, and applications. *The Counseling Psychologist, 34*, 684-718.

Knapp, T. R. (1978). Canonical correlation analysis: A general parametric significance testing system. *Psychological Bulletin, 85*, 410-416.

Loose, F., Régner, I., Morin, A. J. S., & Dumas, F. (in press). Are academic discounting and devaluing double-edged swords? Their relations to global self-esteem, achievement goals, and performance among stigmatized students. *Journal of Educational Psychology.*

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. Mahwah, NJ: Erlbaum.

MacKinnon, D. P., Lockwood, C. M., & Williams, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behavioral Research, 39*, 99-128.

Marsh, H. W. (1983). Multidimensional ratings of teaching effectiveness by students from different academic settings and their relation to student/course/instructor characteristics. *Journal of Educational Psychology, 75,* 150-166.

Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research, 11,* 253-388. (Whole Issue No. 3)

Marsh, H. W. (1991a). A multidimensional perspective on students' evaluations of teaching effectiveness: A reply to Abrami and d'Apollonia (1991). *Journal of Educational Psychology, 83*, 416-421.

Marsh, H. W. (1991b). Multidimensional students' evaluations of teaching effectiveness: A test of alternative higher-order structures. *Journal of Educational Psychology, 83,* 285-296.

Marsh, H. W. (2007). Application of confirmatory factor analysis and structural equation

modeling in sport/exercise psychology. In G. Tenenbaum & R. C. Eklund (Eds.), *Handbook of sport psychology* (3rd ed., pp. 774-798). New York: Wiley.

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., 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., & Cheng, J. H. S (2012). Physical self-concept. In G. Tenenbaum, R. Eklund, & A. Kamata (Eds.), *Measurement in Sport and Exercise Psychology* (pp. 215-226). Champaign, IL: Human Kinetics.

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: A Multidisciplinary Journal, 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., 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. McArdle (Eds.), *Psychometrics. A Festschrift to Roderick P. McDonald* (pp. 275-340). Hillsdale, NJ: Erlbaum.

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

Marsh, H. W., Hau, K.-T., Wen, Z., Nagengast, B., & Morin, A. J. S. (in press). Moderation. In T. D. Little (Ed.), *Oxford handbook of quantitative methods*. New York: Oxford University Press.

Marsh, H. W., & Hocevar, D. (1991). The multidimensionality of students' evaluations of teaching effectiveness: The generality of factor structures across academic discipline, instructor level, and course level. *Teaching and Teacher Education, 7,* 9-18.

Marsh, H. W., Liem, G. A. D., Martin, A. J., Morin, A. J. S., & Nagengast, B. (2011). Methodological-measurement fruitfulness of exploratory structural equation modeling (ESEM): New approaches to key substantive issues in motivation and engagement. *Journal of Psychoeducational Assessment, 29,* 322-346

Marsh, H. W., Lüdtke, O., Muthén, B. O., Asparouhov, T., Morin, A. J .S., & Trautwein, U. (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). *Two wrongs do not make a right: Camouflaging misfit with item-parcels in CFA.* Centre for Positive Psychology and Education, University of Western Sydney, Australia.

Marsh, H. W., Martin, A., & Debus, R. (2001). Individual differences in verbal and math self-perceptions: One factor, two factors, or does it depend on the construct? In R.

Riding & S. Rayner (Eds.). *Self perception: International perspectives on individual differences* (pp. 149-170). Westport, CT: Ablex.

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

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

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.

Marsh, H. W., Nagengast, B., Morin, A. J. S., Parada, R. H., Craven, R. G., & Hamilton, L. R. (2011). Construct validity of the multidimensional structure of bullying and victimization: An application of exploratory structural equation modeling. *Journal of Educational Psychology, 103,* 701-732.

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., & Yeung, A. S. (1998). Top-down, bottom-up, and horizontal models: The direction of causality in multidimensional, hierarchical self-concept models. *Journal of Personality and Social Psychology, 75,* 509-527.

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

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

McDonald, R. P. (1985). *Factor analysis and related methods.* Hillsdale, NJ: Erlbaum.

Meleddu, M., Guicciardi, M., Scalas, L. F., & Fadda, D. (2012). Validation of an Italian version of the Oxford Happiness Inventory in adolescence. *Journal of Personality Assessment, 94,* 175-185.

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

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York, NY: Routledge.

Morin, A. J. S., & Maïano, C. (2011). Cross-validation of the Short Form of the Physical Self-Inventory (PSI-S) using Exploratory Structural Equation Modeling (ESEM). *Psychology of Sport and Exercise, 12,* 540-554.

Morin, A. J. S., Maïano, C., Nagengast, B., Marsh, H. W., Morizot, J., & Janosz, M. (2011). General growth mixture analysis of adolescents' developmental trajectories of anxiety: The impact of untested invariance assumptions on substantive interpretations. *Structural Equation Modeling: A Multidisciplinary Journal, 18,* 613-648.

Muthén, B. O. (1989). Latent variable modeling in heterogenous populations. *Psychometrika, 54,* 557-585.

Muthén, B. O. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika, 29,* 81-117.

Muthén, L. K., & Muthén, B. O. (2009). *Mplus short courses: Topic 1: Exploratory factor analysis, confirmatory factor analysis, and structural equation modeling for continuous outcomes.* Los Angeles CA: Muthén & Muthén. Retrieved from http://www.statmodel.com/course_materials.shtml

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

Myers, N. D., Chase, M. A., Pierce, S. W., & Martin, E. (2011). Coaching efficacy and exploratory structural equation modeling: A substantive-methodological synergy. *Journal of Sport and Exercise Psychology*, *33*, 779-806.

Neale, M. C., Boker, S. M., Xie, G., & Maes, H. H. (2003). *Mx: Statistical modeling*. Richmond: Virginia Commonwealth University, Department of Psychiatry.

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.

Sass, D. A., & Schmitt, T. A. (2010). A comparative investigation of rotation criteria within exploratory factor analysis. *Multivariate Behavioral Research, 45*, 1-33.

Selig, J. P., & Preacher, K. J. (2009). Mediation models for longitudinal data in developmental research. *Research in Human Development, 6*, 144-164.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. New York, NY: Chapman & Hall/CRC.

Spearman, C. (1904). "General intelligence," objectively determined and measured. *American Journal of Psychology, 15*, 201-293.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications.* Washington, DC: American Psychological Association.

Thompson, B., & Daniel, L. G. (1996). Factor analytic evidence for the construct validity of

scores: A historical overview and some guidelines. *Educational and Psychological Measurement, 56,* 197-208.

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

Toland, M. D., & De Ayala, R. J. (2005). A multilevel factor analysis of students' evaluations of teaching. *Educational and Psychological Measurement, 65, 272-296.*

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3,* 4-70.

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.

Yates, A. (1987). *Multivariate exploratory data analysis: A perspective on exploratory factor analysis*. Albany, NY: State University of New York Press.

Table 1. Results from the total group and multiple group cross sectional models

Model	χ^2	df	RMSEA	RMSEA 95% CI	CFI	TLI
Main models						
Time 1 data CFA model (CFA-Time1.inp)	339.062*	8	.096	.087-.105	.968	.940
Time 1 data ESEM model (ESEM-Geo.5-Time1.inp)	3.085	4	.000	.000-.020	1.000	1.000
Time 2 data CFA model (CFA-Time2.inp)	338.377*	8	.096	.087-.105	.969	.941
Time 2 data ESEM model (ESEM-Geo.5-Time2.inp)	7.864	4	.015	.000-.030	1.000	.999
Multiple group invariance models (Time 1 data)						
Model 1: Configural (MG-inv-M1.inp)	9.064	8	.008	.000-.027	1.000	1.000
Model 2: Load. (MG-inv-M2.inp)	11.646	16	.000	.000-.014	1.000	1.001
Model 3: Load., uniq. (MG-inv-M3.inp)	17.730	22	.000	.000-.013	1.000	1.001
Model 4: Load., FVC. (MG-inv-M4.inp)	18.204	19	.000	.000-.018	1.000	1.000
Model 5: Load., int. (MG-inv-M5.inp)	246.682*	20	.071	.063-.079	.976	.964
Model 5p: Load., p.int. (MG-inv-M5p.inp)	16.163	19	.000	.000-.015	1.000	1.000
Model 6: Load., uniq., FVC. (MG-inv-M6.inp)	24.125	25	.000	.000-.016	1.000	1.000
Model 7: Load., int., uniq. (MG-inv-M7.inp)	256.920*	26	.063	.056-.070	.976	.972
Model 7p: Load., p.int., uniq. (MG-inv-M7p.inp)	22.342	25	.000	.000-.015	1.000	1.000
Model 8: Load., int., FVC. (MG-inv-M8.inp)	253.330*	23	.067	.059-.074	.976	.969
Model 8p: Load., p.int., FVC. (MG-inv-M8p.inp)	22.721	22	.004	.000-.018	1.000	1.000
Model 9: Load., int., uniq., FVC. (MG-inv-M9.inp)	263.316*	29	.060	.053-.067	.976	.975
Model 9p: Load., p.int., uniq., FVC. (MG-inv-M9p.inp)	28.732	28	.003	.000-.017	1.000	1.000
Model 10: Load., int., FMeans. (MG-inv-M10.inp)	641.270*	22	.112	.104-.119	.935	.912
Model 10p: Load., p.int., FMeans. (MG-inv-M10p.inp)	319.936*	21	.080	.072-.087	.969	.955
Model 11: Load., int., uniq., FMeans. (MG-inv-M11.inp)	654.880*	28	.100	.093-.106	.934	.930
Model 11p: Load., p.int., uniq., FMeans. (MG-inv-M11p.inp)	326.470*	27	.070	.063-.077	.969	.965
Model 12: Load., int., FVC., FMeans. (MG-inv-M12.inp)	641.699*	25	.105	.098-.112	.936	.923
Model 12p: Load., p.int., FVC., FMeans. (MG-inv-M12p.inp)	320.123*	24	.074	.067-.081	.969	.961
Model 13: Load., int., uniq., FVC., FMeans. (MG-inv-M13.inp)	655.453*	31	.095	.088-.101	.935	.937
Model 13p: Load., p.int., uniq., FVC., FMeans. (MG-inv-M13p.inp)	326.628*	30	.066	.060-.073	.969	.969
MIMIC models						
Time 1 MIMIC null effect (MIMIC-null-Time1.inp)	634.162*	10	.118	.110-.126	.943	.881
Time 1 MIMIC saturated (MIMIC-satur-Time1.inp)	2.833	4	.000	.000-.019	1.000	1.001
Time 1 MIMIC invariant intercept (MIMIC-base-Time1.inp)	242.024*	8	.081	.072-.090	.979	.944
Time 1 MIMIC partially invariant intercept (MIMIC-DIF-Time1.inp)	7.440	7	.004	.000-.019	1.000	1.000
Time 2 MIMIC null effect (MIMIC-null-Time2.inp)	729.323*	10	.126	.119-.134	.936	.865
Time 2 MIMIC saturated (MIMIC-satur-Time2.inp)	7.785	4	.015	.000-.030	1.000	.998
Time 2 MIMIC invariant intercept (MIMIC-base-Time2.inp)	322.398*	8	.093	.085-.102	.972	.926
Time 2 MIMIC partially intercept invariant (MIMIC-DIF-Time2.inp)	8.655	7	.007	.000-.021	1.000	1.000

Note. Names of the input file in the supplementary materials are reported in parentheses; *: $p < .05$; CFA: Confirmatory factor analysis; ESEM: Exploratory Structural Equation Modeling; χ^2: Chi square test of model fit; df: degrees of freedom; RMSEA: Root Mean Square Error of Approximation; RMSEA 95% CI: 95% confidence interval of the RMSEA; TLI: Tucker-Lewis Index; CFI: Comparative Fit Index; Load: Loadings invariance; Uniq: Uniquenesses invariance; FVC: Factor variance-covariance invariance; Int: Intercepts invariance; FMeans: Factor means invariance.

Table 2. Standardized parameters from the CFA and ESEM models based on Time 1 data.

		CFA			ESEM geomin, ε = .5	
Item	**F1**	**F2**	**Uniq.**	**F1**	**F2**	**Uniq.**
Time 1		(CFA-Time1.inp)			(ESEM-Geo.5-Time1.inp)	
X1	.722		.478	.833	-.038	.333
X2	.843		.289	.690	.232	.330
X3	.742		.450	.536	.318	.461
X4		.835	.303	.139	.766	.300
X5		.774	.401	.166	.674	.420
X6		.510	.739	-.046	.566	.700
Correlations	.728			.439		
Time 2		(CFA-Time2.inp)			(ESEM-Geo.5-Time2.inp)	
Y1	.737		.457	.837	-.032	.322
Y2	.841		.292	.698	.222	.330
Y3	.751		.436	.560	.303	.449
Y4		.846	.284	.149	.762	.300
Y5		.776	.398	.140	.704	.400
Y6		.492	.758	-.054	.551	.719
Correlations	.709			.429		

Note. Names of the input file in the supplementary materials are reported in parentheses; All coefficients significant at the .05 level; CFA: Confirmatory factor analysis; ESEM: Exploratory Structural Equation Modeling; F1: standardized loadings on the first factor; F2: standardized loadings on the second factor; Uniq.: standardized uniquenesses.

Table 3. Marsh et al. (2009, 2010) Taxonomy of Invariance Tests

Model	Invariant parameters	Signification	Nesting
Model 1	None, apart from the number of factors.	Configural Invariance	none
Model 2	Factor loadings.	Weak measurement invariance	1
Model 3	Factor loadings, items' uniquenesses.		1, 2
Model 4	Factor loadings, factor variances-covariance.		1,2
Model 5	Factor loadings, items' intercepts.	Strong measurement invariance	1,2
Model 6	Factor loadings, items' uniquenesses, factor variances-covariance.		1,2,3,4
Model 7	Factor loadings, items' intercepts, items' uniquenesses.	Strict measurement invariance	1,2,3,5
Model 8	Factor loadings, items' intercepts, factor variances-covariance.		1,2,4,5
Model 9	Factor loadings, items' intercepts, items' uniquenesses, factor variances-covariance.		1-8
Model 10	Factor loadings, items' intercepts, factor means.	Latent mean invariance	1,2,5
Model 11	Factor loadings, items' intercepts, items' uniquenesses, factor means.	Manifest mean invariance	1,2,3,5,7,10
Model 12	Factor loadings, items' intercepts, factor variances-covariance, factor means.		1,2,4,5,6,10
Model 13	Factor loadings, items' intercepts, items' uniquenesses, factor variances-covariance, factor means.	Complete factorial invariance	1-12

Models with latent factor means freely estimated constrain intercepts to be invariant across groups, while models where intercepts are free imply that mean differences are a function of intercept differences. Nesting relations are shown such that 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) while model 13 (complete invariance) is nested under all other models.
Partially adapted from Table 1 of Marsh, Muthén, Asparouhov, Lüdtke, Robitzsch, Morin, & Trautwein (2009).

Table 4. Results from the longitudinal models

Model	χ^2	df	RMSEA	RMSEA 95% CI	CFI	TLI
Main models						
CFA model without CUs. (Longit-CFA.inp)	2605.798*	48	.109	.105-.112	.896	.857
CFA model with CUs. (Longit-CFA-CU.inp)	679.461*	42	.058	.054-.062	.974	.959
ESEM model without CUs. (Longit-ESEM.inp)	2002.267*	40	.104	.101-.108	.920	.869
ESEM model with CUs. (Longit-ESEM-CU.inp)	62.474*	34	.014	.008-.019	.999	.998
Longitudinal invariance models (with CUs)						
Model 1: Configural (Long-inv-M1.inp)	62.474 *	34	.014	.008-.019	.999	.998
Model 2: Load. (Long-inv-M2.inp)	70.407*	42	.012	.007-.017	.999	.998
Model 3: Load., uniq. (Long-inv-M3.inp)	73.458*	48	.011	.005-.016	.999	.999
Model 4: Load., FVC. (Long-inv-M4.inp)	76.268*	45	.012	.007-.017	.999	.998
Model 5: Load., int. (Long-inv-M5.inp)	474.364*	46	.045	.042-.049	.983	.975
Model 5p: Load., p.int. (Long-inv-M5p.inp)	74.521*	45	.012	.007-.017	.999	.998
Model 6: Load., uniq., FVC. (Long-inv-M6.inp)	80.125*	51	.011	.006-.016	.999	.998
Model 7: Load., int., uniq. (Long-inv-M7.inp)	477.276*	52	.043	.039-.046	.983	.978
Model 7p: Load., p.int., uniq. (Long-inv-M7p.inp)	77.565*	51	.011	.005-.015	.999	.999
Model 8: Load., int., FVC. (Long-inv-M8.inp)	480.265*	49	.044	.041-.048	.983	.976
Model 8p: Load., p.int., FVC. (Long-inv-M8p.inp)	80.403*	48	.012	.007-.017	.999	.998
Model 9: Load., int., uniq., FVC. (Long-inv-M9.inp)	483.955*	55	.042	.038-.045	.983	.979
Model 9p: Load., p.int., uniq., FVC. (Long-inv-M9p.inp)	84.264*	54	.011	.006-.016	.999	.998
Model 10: Load., int., FMeans. (Long-inv-M10.inp)	1481.685*	48	.081	.078-.085	.942	.920
Model 10p: Load., p.int., FMeans. (Long-inv-M10p.inp)	1119.971*	47	.071	.068-.075	.956	.939
Model 11: Load., int., uniq., FMeans. (Long-inv-M11.inp)	1484.599*	54	.077	.073-.080	.942	.929
Model 11p: Load., p.int., uniq., FMeans. (Long-inv-M11p.inp)	1122.936*	53	.067	.064-.070	.957	.946
Model 12: Load., int., FVC., FMeans. (Long-inv-M12.inp)	1488.286*	51	.079	.076-.083	.942	.925
Model 12p: Load., p.int., FVC., FMeans. (Long-inv-M12p.inp)	1126.449*	50	.069	.066-.073	.956	.942
Model 13: Load., int., uniq., FVC., FMeans. (Long-inv-M13.inp)	1491.877*	57	.075	.072-.078	.942	.933
Model 13p: Load., p.int., uniq., FVC., FMeans. (Long-inv-M13p.inp)	1130.176*	56	.065	.062-.069	.956	.949

Note. Names of the input file in the supplementary materials are reported in parentheses; *: $p \leq .05$; CFA: Confirmatory factor analysis; ESEM: Exploratory Structural Equation Modeling; χ^2: Chi square test of model fit; df: degrees of freedom; RMSEA: Root Mean Square Error of Approximation; RMSEA 95% CI: 95% confidence interval of the RMSEA; TLI: Tucker-Lewis Index; CFI: Comparative Fit Index; Load: Loadings invariance; Uniq: Uniquenesses invariance; FVC: Factor variance-covariance invariance; Int: Intercepts invariance; FMeans: Factor means invariance.

Table 5. Correlations between factors in the longitudinal models

	CFA models			ESEM models		
	Time 1 Factor 1	Time 1 Factor 2	Time 2 Factor 1	Time 1 Factor 1	Time 1 Factor 2	Time 2 Factor 1
Models without CUs.						
Time 1 Factor 2	.728			.447		
Time 2 Factor 1	.545	.307		.559	.143	
Time 2 Factor 2	.342	.530	.709	.195	.551	.439
Models with CUs.						
Time 1 Factor 2	.728			.438		
Time 2 Factor 1	.489	.304		.495	.166	
Time 2 Factor 2	.338	.432	.708	.215	.438	.430

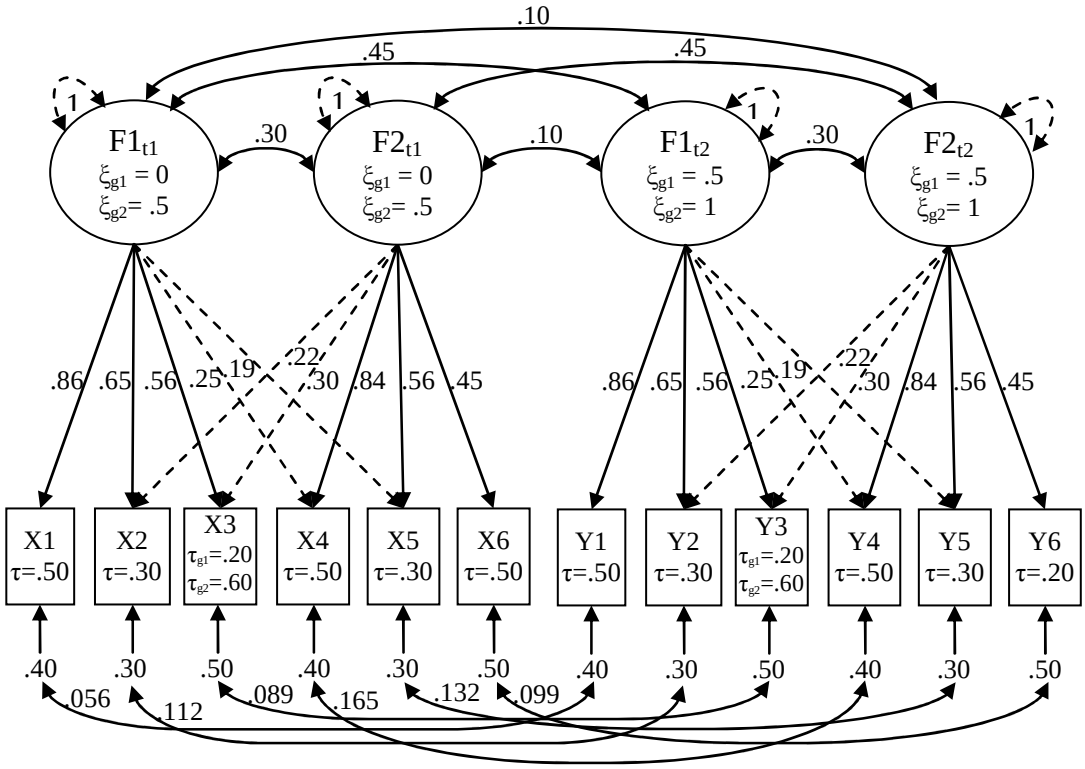

Figure 1. Graphical representation of the population generating model.

Note. Ovals represent latent factors and squares represent observed variables; full unidirectional arrows linking ovals and squares represent the main factor loadings; dotted unidirectional arrows linking ovals and squares represent the cross-loadings; full unidirectional arrows placed under the squares represent the item uniquenesses; bidirectional full arrows linking the ovals represent factor covariances/correlations; bidirectional full arrows linking the squares represent the longitudinal correlated uniquenesses; bidirectional dashed arrows connecting a single oval represent factor variances; τ represents items intercepts; ξ represents the latent factor means; the subscripts g1 and g2 indicate that a parameter varies across both simulated subgroups.

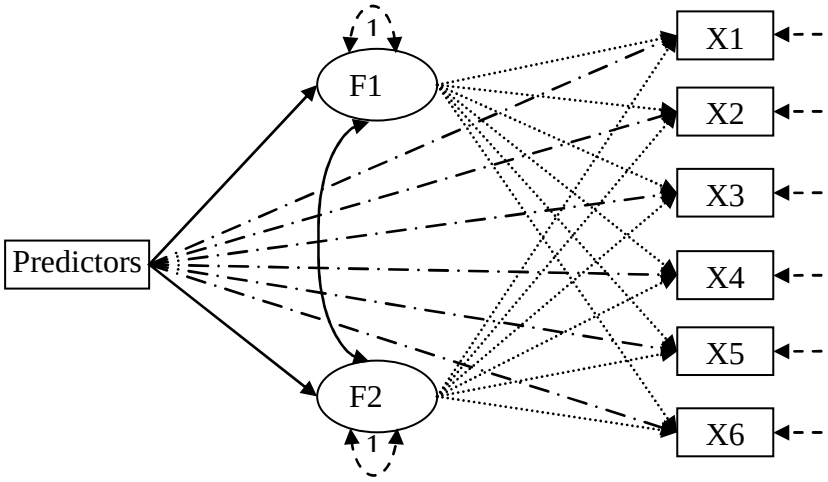

Figure 2. MIMIC model used to test differential item functioning.

Note. Ovals represent latent factors and squares represent observed variables; dotted unidirectional arrows represent factor loadings; dashed unidirectional arrows represent items uniquenesses; full unidirectional arrows represent the paths needed to estimate the MIMIC model with invariant intercepts; dashed-and-dotted unidirectional arrows represent the paths needed to estimate the saturated MIMIC model; bidirectional full arrows linking the ovals represent factor covariances/correlations; bidirectional dashed arrows connecting a single oval represent factor variances.

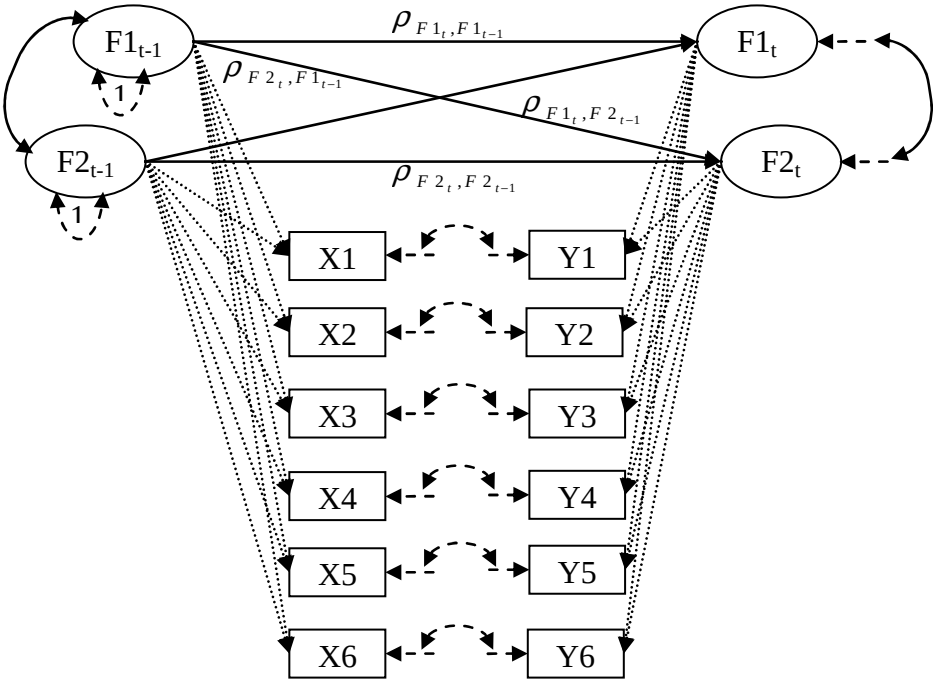

Figure 3. Autoregressive cross lagged ESEM model.

Note. Ovals represent latent factors and squares represent observed variables; dotted unidirectional arrows represent factor loadings; dashed unidirectional arrows represent items uniquenesses and factor disturbances; full unidirectional arrows represent the autoregressive and cross lagged predictive paths; bidirectional full arrows linking the ovals represent time-specific factor covariances/correlations; bidirectional dashed arrows connecting a single oval represent factor variances; bidirectional dashed arrows linking the squares represent the longitudinal correlated uniquenesses.

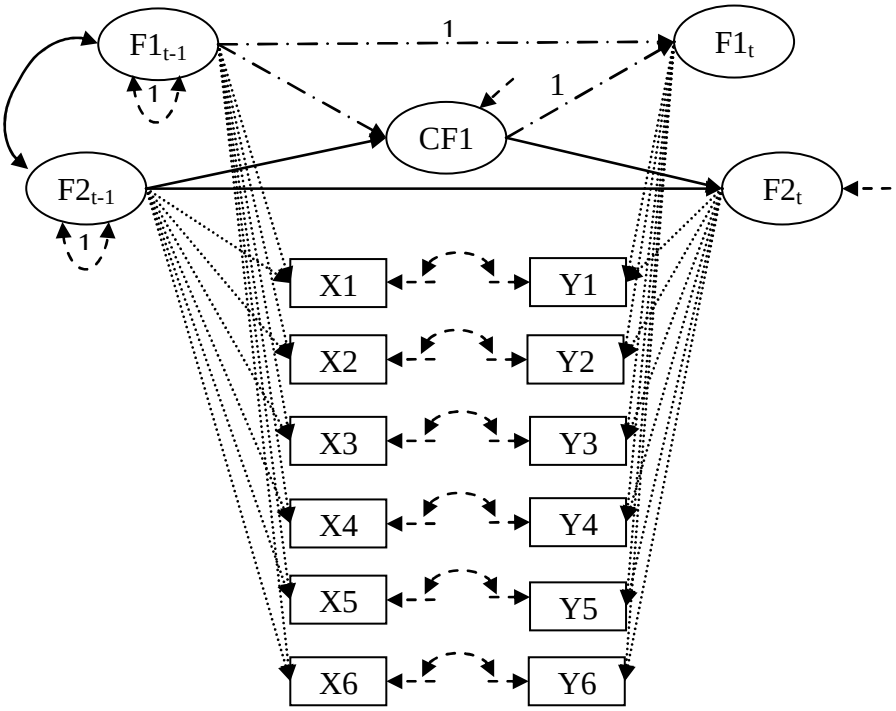

Figure 4. Longitudinal mediation EWC model with latent difference score for factor 1.
Note. Ovals represent latent factors and squares represent observed variables; dotted unidirectional arrows represent factor loadings; dashed unidirectional arrows represent items uniquenesses and factor disturbances; full unidirectional arrows represent the longitudinal mediation paths; dashed-and-dotted unidirectional arrows are needed to estimate the CF1 latent difference variable; bidirectional full arrows linking the ovals represent time-specific factor covariances/correlations; bidirectional dashed arrows connecting a single oval represent factor variances; bidirectional dashed arrows linking the squares represent the longitudinal correlated uniquenesses.