Exploratory Structural Eqation Modeling: An Integration of the Best Features of Exploratory and Confirmatory Factor Analysis

Herbert W. Marsh; University of Western Sydney, Australia, University of Oxford, UK, & King Saud University, Saudi Arabia

Alexandre J. S. Morin; University of Western Sydney, Australia

Philip D. Parker; University of Western Sydney, Australia

Gurvinder Kaur, University of Western Sydney, Australia

Corresponding author

Requests for further information about this investigation should be sent to Professor Herbert W. Marsh, Centre for Positive Psychology and Education, University of Western Sydney; E-mail: *h.marsh@uws.edu.au*.

Supplemental Material: Annu. Rev. Clin. Psychol. 2014. 10:85-110
doi: 10.1146/annurev-clinpsy-032813-153700
Exploratory Structural Equation Modeling: An Integration of the Best Features of Exploratory and Confirmatory Factor Analysis
Marsh et al.

SUPPLEMENTAL MATERIALS

Formal Statistical Basis of ESEM

In the ESEM model (Asparouhov & Muthén, 2009; Marsh et al., 2009; for more technical detail, also see http://www.statmodel.com/esem.shtml), there are *p* dependent variables Y = (Y1, ..., Yp), *q* independent variables X = (X1, ..., Xq), and *m* latent variables $\eta = (\eta_1, ..., \eta_m)$:

$$\mathbf{Y} = \nu + \Lambda\eta + K\mathbf{X} + \varepsilon \qquad [1]$$

$$\eta = \alpha + B\eta + \Gamma\mathbf{X} + \zeta . \qquad [2]$$

Standard assumptions of this model are that the ε and ζ residuals are normally distributed with mean 0 and variance covariance matrix θ and ψ respectively. The first equation represents the measurement model where ν is a vector of intercepts, Λ is a factor loading matrix, η is a vector of continuous latent variables, **K** is a matrix of **Y** on **X** regression coefficients, and ε is a vector of residuals for **Y**. The second equation represents the latent variable model where α is a vector of latent intercepts, **B** is a matrix of η on η regression coefficients, Γ is a matrix of η on **X** regression coefficients, and ζ is a vector of latent variable residuals.

In ESEM, η can include multiple sets of ESEM factors defined either as ESEM or CFA factors (noting that in the most basic model with no further constraints, ESEM factors are merely traditional EFA factors). More precisely, the CFA factors are identified as in traditional CFA/SEMs, where each factor is associated with a different set of indicators. ESEM factors can be divided into blocks of factors so that a series of indicators is used to estimate all ESEM factors within a single block, and a different set of indicators is used to estimate another block of ESEM factors. However, specific items may be assigned to more than one set of ESEM or CFA factors. The assignment of items to CFA and/or ESEM factors is usually determined on the basis of a priori theoretical expectations, practical considerations, or, perhaps, post-hoc, based on preliminary tests conducted on the data. The integrative framework provided by ESEM is demonstrated, in that ESEM is appropriate in any combination of ESEM and CFA factors, and is easily extended to accommodate SEMs involving ESEM or CFA factors.

Goodness of fit

In applied CFA/SEM research, applied researchers seek 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 indexes 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 researchers still treat the indexes and recommended cut-offs as golden rules. Others argue that fit indexes should be discarded altogether, while a few argue that we should dispense with multiple indicators altogether and rely solely on chi-square goodness-of-fit indexes, and many (like us) argue that fit indices should 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, small deviations from multivariate normality, and minor misspecifications, applied CFA/SEM research generally focuses on indices that are sample size independent (Hu & Bentler 1999, Marsh Balla & Hau 1996; Marsh Hau & Wen 2004, Marsh Hau 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 population values of TLI and CFI vary along a 0-to-1 continuum, and values greater than .90 and .95 typically reflect acceptable and excellent fits respectively to the data. Values smaller than .08 or .06 for the RMSEA support acceptable and good model fits respectively.

In most studies, it is useful to compare the fit of alternative, competing models. In this case, comparison of the relative fit of models imposing more or fewer invariance constraints is more important than the absolute level of fit for any one model—so long as the fit of the bestfitting model is acceptable. Thus, for example, the ESEM model with cross-loadings is nested under corresponding CFA/SEM with an ICM-CFA structure. Although the chi-square difference test can be used for this purpose, this test suffers even more problems than the chi- square test for single models that led to the development of fit indexes (see Marsh et al. 1998). Cheung and Rensvold (2001) and Chen (2007) suggest that if the decrease in fit for the more parsimonious model is less than .01 for incremental fit indices like the CFI, there is reasonable support for the more parsimonious model. Chen (2007) suggests 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; there is considerable evidence that realistically large factor structures (e.g., instruments with at least 50 items and at least 5 factors) are typically unable to satisfy even the minimally acceptable standards of fit (Marsh 2007, Marsh Hau et al. 2005).

More research is needed on the appropriateness of traditional indices and approaches to evaluating model fit for ESEM studies. Given the lack of consensus about fit indexes, it is not surprising that there is also ambiguity about their application in ESEM and in regard to the new issues that ESEM raises. For example, because the number of factor loadings alone in ESEM applications is the product of the number of items times the number of factors, the total number of parameter estimates in ESEM applications can be massively more than in CFA. This feature might make problematic any index that does not control for parsimony (due to capitalization on chance) and call into question the appropriateness of controls for parsimony in indexes that do. In the meantime, we suggest that applied researchers use an eclectic approach based on the subjective integration of a variety of different indexes, detailed evaluations of the actual parameter estimates in relation to theory, a priori predictions, common sense, and comparison of viable alternative models specifically designed to evaluate goodness of fit in relation to key issues.

Complex measurement error structures

The ability to model complex error structures is an important advantage that ESEM shares with CFA, and that distinguishes it from traditional approaches to EFA. Although typically it is dubious to introduce correlated uniquenesses ex post facto to improve fit, there are numerous situations in which models with a priori CUs should be the default model. Thus, for example, when the same measures are collected on multiple occasions, the a priori model should include CUs for responses to the same items on different occasions. Similarly, when different factors in a multidimensional instrument are based on parallel worded items, the a priori should also include CUs relating items with parallel wording. In models of complex multitrait-multimethod (MTMM) data structures, a particularly useful model represents method effects as CUs between indicators representing the method (e.g., those from the same instrument if the multiple instruments are the multiple methods, those from the same informant if the multiple informants are the multiple methods). Of course, it is always reasonable to compare these a priori CUs with a more parsimonious model with no CUs. Applied researchers might also posit a priori error structures to account for method effects. Importantly, ESEM is similar to CFA/SEM, in terms of this added flexibility that is not available with traditional approaches to EFA (see subsequent discussion).

MIMIC Models of Relations with Background and Predictor Variables

The MIMIC model (Jöreskog & Goldberger 1975, Marsh Ellis et al. 2005, Marsh Tracey & Craven 2006, Muthén 1989) is a multivariate regression model in which latent variables are regressed on predictors. The MIMIC model has *important* advantages over the multigroup 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 model to be estimated in each group. Also, it allows researchers to consider multiple independent variables that would typically become unmanageable in multigroup analyses. The multigroup approach (like a traditional ANOVA with fully manifest scores) is limited to situations in which the independent MIMIC variable can be represented by a small number of discrete categories. Although some variables (e.g., gender, diagnostic categories) are naturally categorical, many (e.g., age, pre-test scores) are not. In psychological research, it is well known that there are serious limitations in using a small number of categories to represent a reasonably continuous variable like age (MacCallum et al. 2002), particularly when the continuous predictor variable might have non-linear effects. Importantly, the MIMIC model (like the multiple regression extension of ANOVA with fully manifest scores) is equally appropriate for categorical variables (e.g., gender, diagnostic categories), continuous manifest predictors (e.g. age, income, single-item pretest scores), continuous latent factors based on multiple indicators (e.g. background or pretest constructs based on multiple indicators), or a mixture of these different types of background predictor variables.

The MIMIC approach is easily extended to include interactive and non-linear effects. In clinical psychology, for example, it is often hypothesized that the effect of an intervention will interact with characteristics of individual participants (e.g., a social skills training program developed for shy individuals might not be effective for extraverted individuals). More generally, such aptitude-treatment interactions are posited in many areas of psychology, and many psychological theories explicitly hypothesize interaction effects. Analyses of quadratic effects are also important. Thus, for example, low levels of anxiety might facilitate performance, whereas high levels of anxiety might undermine performance; intervention strength (or dosage level) and outcome variables might have non-linear relations, such that benefits increase up to an optimal level, and then level off or even decrease beyond this optimal point. Intervention effects might be expected to have non-linear relations with time, to be the highest immediately following the intervention, and to decrease in a nonlinear manner over time following the intervention.

A particular strength of the MIMIC approach is that it is readily extended to more complex models of linear and non-linear components of individual MIMIC variables, to mediation involving additive effects of multiple variables, or moderation (interaction) between multiple variables. Baron and Kenny's (1986) classic distinction between moderation and mediation is important in understanding the MIMIC approach. Mediation occurs when some of the effects of an independent variable (X) on the dependent variable (Y) can be explained in terms of another mediating variable (MED) that falls between X and Y in terms of causal ordering. Thus, for example, the effects of family socioeconomic status (SES) at the start of high school (X) on achievement test scores at the end of high school (Y) are likely to be mediated in part by coursework during high school (MED). Critical assumptions in most mediation models are a strict causal ordering of X➔Med➔Y, but this assumption is usually very difficult to test and is often largely ignored in mediation studies. In contrast, moderation is said to have taken place when the size or direction of the effect of X on Y varies with the level of a moderating variable (MOD). Thus, for example, the effects of a remedial course instead of regular coursework (X) on subsequent achievement (Y) may vary systematically, depending on the student's initial level of ability (MOD); the effect of the remedial course may be very positive for initially less able students, negligible for average-ability students, and even detrimental for high-ability students who would probably gain more from regular or advanced mathematics coursework. Of course, mediation, moderation, and non-linear effects are not mutually exclusive and it is possible—even desirable—to incorporate all three into the same latent variable models (e.g., Nagengast & Marsh, 2012).

Marsh Wen et al. (2013) provide an overview of methods for the analysis of products between latent variables—latent interactions and latent quadratic effects. If both independent and dependent variables are manifest, it might be appropriate to use ANOVA (if both independent variables are categorical) or the multiple regression approach to ANOVA (if at least one of the independent variables is reasonably continuous). When one of the independent variables is a categorical variable that can be represented by a relatively small number of groups, the multiple-group CFA/SEM or ESEM approach is an attractive alternative. However, for fully latent models where both independent and dependent variables are latent, Marsh Wen et al. (2013) review a variety of new and evolving approaches to evaluate latent product models that have important advantages over traditional manifest approaches. In particular, these approaches allow researchers to incorporate the measurement model into structural tests of the product terms—providing strong tests of the a priori measurement model, controlling for complex sorts of measurement error that attenuate effects of product terms, and providing much greater flexibility in the nature of the models considered. Although all these—and other— advantages of latent variable models over manifest models are relevant, they are particularly important to the evaluation of product models that depend so strongly on the underlying measurement of the interacting constructs and on controlling for measurement error.

However, the major limitation of the MIMIC model is that it assumes that latent outcome variables are invariant over different levels of the MIMIC variable. Furthermore, these critical assumptions are not easily tested within the MIMIC framework. Thus, for example, if gender is related to the latent factor of depression, it is not easy to test whether the factor loadings relating the depression items to the latent construct are invariant over gender—an implicit, typically untested, assumption in the application of the MIMIC approach. We return to this issue in the next section, where we discuss the issues of measurement invariance in more detail, and a hybrid approach integrating the MIMIC and multigroup approaches. We also note that the MIMIC model can be applied to either ESEM or CFA factors, and that the decision as to which is most appropriate should be based upon a comparison of the two different models. If the MIMIC model with a highly restrictive ICM-CFA factor structure does not adequately fit the data, where the ESEM MIMIC model does, the ESEM MIMIC model probably should be used instead. However, the ESEM MIMIC model clearly reflects the flexibility of the ESEM approach in relation to traditional EFAs.

Models of Causal ordering

Studies of causal ordering testing a priori hypotheses about the direction of causation have important relevance to clinical research (e.g., does depression lead to anxiety, does anxiety lead to depression, or are the two constructs reciprocally related?). These questions can be investigated in the context of autoregressive cross-lagged models (Jöreskog 1979, 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. There is a well-established paradigm, based on SEMs, for evaluating the directionality of the associations when the same constructs are measured on multiple occasions (e.g., Marsh & Craven 2006). Extending this approach, separate estimates were obtained for direct effects (the typical path coefficient), indirect effects mediated through intervening variables, and total effects, the sum of direct and indirect effects (see related discussion by Little et al. 2007, Marsh & O'Mara 2010). Although these models have been routinely studied with CFA/SEMs, Marsh, Nagengast et al. (2011) have illustrated how these questions can be addressed with ESEM. Clearly this ESEM approach to models of causal ordering demonstrates the flexibility of ESEM compared to traditional EFAs. The approaches based on ESEM and on CFA/SEMs are similar. However, unless the highly restrictive IMC-CFA factor structure is able to fit the data, biased estimates based on the misspecified models are likely to bias structural parameters in ways that are not easily predicted. Again, it is important to reiterate that the highly restrictive autoregressive cross- lagged models based on CFA/SEM are merely a special case of and nested under, the more general ESEM. When the differences between the two are non-significant, small or substantively unimportant, the CFA/SEM approach is preferred on the basis of parsimony, but a growing body of experience suggests that the fit of ESEMs is meaningfully better, so that interpretations of structural parameters are likely to be more valid.

Measurement invariance

Of particular substantive importance for clinical research are mean-level differences across multiple groups (e.g., male vs. female; age groups; clinical vs. non-clinical populations; treatment vs. control groups) or over time (i.e., observing the same group of participants at multiple occasions, perhaps before and after an intervention). Typically, tests of whether the underlying factor structure is the same for different groups or multiple occasions have been ignored in clinical research. However, these mean comparisons assume the invariance of at least factor loadings and item intercepts (problems associated with differential item functioning). 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 are potentially invalid. For example, if gender or longitudinal differences vary substantially for different items used to infer a construct in a manner that is unrelated to respondents' true levels on the latent construct, then the observed differences might be idiosyncratic to the particular items used. From this perspective, it is important to be able to evaluate the full measurement invariance of responses.

Tests of measurement invariance evaluate the extent to which measurement properties generalize over multiple groups, situations, or occasions (Meredith 1993, Vandenberg & Lance 2000). Measurement invariance is fundamental to the evaluation of construct validity and generalizability, and an important prerequisite to any form of valid group-based comparison. Historically, multigroup tests of invariance were seen as a fundamental advantage of CFA/SEM over EFA approaches that were largely limited to descriptive comparisons of the factor loadings estimated separately in each group (but see Dolan et al. for an EFA precursor to the more general ESEM framework).

In contrast to traditional EFAs, but like CFAs, ESEMs are easily extended to multigroup tests of invariance. Marsh et al. (2009) operationalized a taxonomy of 13 ESEM models (see Table 1) designed to test measurement invariance that integrates traditional CFA approaches to factor invariance (e.g., Jöreskog & Sörbom 1993, Marsh 1994, 2007; Marsh & Grayson 1994) and IRT approaches to measurement invariance (e.g., Meredith 1964, 1993; also see Millsap 2011, Vandenberg & Lance 2000). Although essentially parallel, these two traditions differ in emphasis. Because the multigroup IRT approach traditionally focused on tests of the unidimensionality of a single construct, relations among different factors (and related issues of cross-loading and complex error structures involving multiple factors) were not given much attention. Because the multigroup CFA approach was traditionally based on covariance matrices that excluded item means, intercept invariance was not given much attention. Integrating these two traditions, Marsh et al. (2009) proposed a 13-model taxonomy of partially nested models to test full measurement invariance and latent means, and applied these models in several ESEM studies. Key models included:

Configural invariance (Model 1). In Model 1—like the traditional EFA approach—no parameter estimates are constrained to be invariant across the groups, although the number of factors is typically constrained to be the same across groups. 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).

Weak measurement invariance (Model 2). Model 2 tests whether the factor loadings are the same across groups or occasions. These are particularly important, both for relating factors to other constructs for different groups with cross-sectional data, and for evaluating patterns of relations among variables in the same group over time with longitudinal data. Also, 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. Such tests of invariance might also be on the basis of selecting items to be retained in early stages of instrument development. Unless the factor loadings are reasonably invariant over groups, any comparisons must be considered suspect, as the constructs themselves differ (i.e., the apples and oranges problem).

Strong measurement invariance (Model 5). Model 5 requires that the indicator intercepts—in addition to the factor loadings—are invariant over groups. Intercept invariance is an important assumption in the comparison of group means across multiple groups (or over time), a violation of which is also called differential item functioning (DIF; technically, monotonic DIF). For example, assume for six items designed to measure a particular trait that three clearly favor one group and three clearly favor the second group. These results provide no basis for evaluating mean differences on the factor, in that even the direction of differences would depend on the particular items used to measure the trait. Furthermore, because these 6 items are only a small sample of items that could be used to evaluate this trait, the results provide only a weak basis for knowing what would happen if a larger, more diverse sample of items were sampled. Support for the invariance of item intercepts would mean that differences based on each of the items considered separately are reasonably consistent in terms of magnitude as well as direction. These results would provide a stronger basis of support for the generalizability of the interpretation of the observed mean differences. Although issues of non- invariance of item intercepts and differential item functioning are well known, historically these issues have been used more frequently in the evaluation of standardized achievement tests than for measures of psycho-social variables and have been largely ignored in traditional EFA approaches.

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.

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 scale or factor scores but is also relevant and interesting in its own right as a test of the generalizability of measurement error across groups. Indeed, the presence of differences in reliability (as represented or absorbed in the item uniquenesses) across the multiple groups could distort mean differences on the observed scores. However, for comparisons based on latent constructs that are corrected for measurement error, the valid comparison of latent means only requires support for strong measurement invariance and not for the additional assumption of the invariance of measurement error. Hence, comparison of group mean differences based on latent-variable models like those considered here makes fewer assumptions than those based on manifest scores. Although the crucial comparison for strict measurement invariance is between Model 5 and Model 7, the invariance of item uniquenesses can also be tested by the comparison of any pair of models differing only in regard to uniquenesses' invariance.

Factor variance-covariance invariance (Model 4). Model 4 tests the invariance of the factor variance-covariance matrix—in addition to factor loadings. Tests of the invariance of the factor variance-covariance matrix (Model 4) are highly relevant to many substantive CFA or SEM studies, where the main focus is on relations among variables. Indeed, such tests have been a major focus of the traditional CFA approach to factorial invariance (e.g., Joreskog 1971, 1979). However, factor variance-covariance invariance is not a necessary precondition in the comparison of latent means, and for this reason has been largely ignored in the traditional IRT approach to measurement invariance. 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).

Invariance of latent means (Models 10 through 13). The final four models (Models 10–

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 be support for the invariance of factor loadings and item intercepts, but not for the invariance of item uniquenesses or the factor variance-covariance matrix. However, typically it is advisable to test the most constrained model that is justified by earlier tests.

Partial Invariance. Although these tests, in each of the 13 models, posit full invariance of all parameter estimates for all groups, Byrne, Shavelson and Muthén (1989, also see Marsh 2007) have argued for the usefulness of a less demanding test of partial invariance in which a subset of parameters are not constrained to be invariant. In particular, partial invariance might be warranted if the sample size is large, there is a sufficient number of items for each factor, and a given set of constraints (e.g., factor loadings, item intercepts, or item uniquenesses) is supported for most of the items for each factor. Also, when there are more than two groups, partial invariance could be limited to a subset of groups. We also note that while tests of partial invariance of factor loadings are not easily applied in ESEM as currently operationalized, new approaches apparently circumvent this problem (e.g., EwC, discussed earlier in printed article and below). 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 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.

Thus far, we have illustrated tests of invariance across multiple groups on a single occasion. However, 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, 1979; Marsh & Hau, 1996; Marsh, 2007). ESEM tests of invariance over time—including the CUs—further demonstrate its flexibility.

Simulated Data Examples (see https://github.com/pdparker/ESEM)

Because of space limitations we were not able to include the details of worked examples (i.e., data, syntax, and discussion of results) in the paper itself, but we have created a separate website with expanded discussion of a data set simulated to reflect a typical clinical application (Morin et al. 2013, https://github.com/pdparker/ESEM). This data set is made available to readers willing to try their hands at ESEM, together with all of the input codes used in this specific illustration. This data set includes six items (X1 through X6) serving as indicators of two correlated factors (Factor 1 and Factor 2), 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. For purposes of illustration, we label these factors Anxiety (Factor 1) and Depression (Factor 2): two correlated/comorbid clinical states which can be measured by indicators/symptoms that can realistically be expected to present significant, and even reasonably large, cross-loadings (for instance, levels on the depressive symptoms of "psychomotor agitation" or "insomnia" can be expected to be elevated in anxious individuals also). For further illustrative purposes, let's imagine that these data were collected as part of a clinical Pre-Test Post-Test design with randomized experimental and control groups. In this context, the first set of items (X1 through X6) will be referred to as Pre- Test data. We also simulated a second set of items (Y1 through Y6), referred to as Post-Test data, designed to represent a second measurement point for the X items and simulated to have similar properties to the Pre-Test data. Two subgroups were simulated to reflect an experimental and a control group. In line with clinical trials, these subgroups were simulated based on the assumption of random assignments (with Pre-Test data showing fully equivalent properties between the groups) and moderately small sample sizes of *n* = 150 for each group (*n* = 300 in total). Further, consistent with common observations in the context of clinical trials, the control group was simulated to show a small decrease of depressive symptoms over time but no change in anxiety level. The experimental group was simulated as showing a substantial decrease in depressive symptoms over time, and with gender-differentiated effects regarding the response to treatment for symptoms of anxiety (see below), consistent with significant construct-specific intervention effects. We simulated the data to show increased levels of measurement error, associated with the last two items of the depression factor in the experimental group, which reflect the occurrence of some disturbance (e.g., noise, fire alarm) toward the end of the testing session in one of the groups.

To add some complexity, we further simulated slightly different models for subgroups reflecting male (*n* = 150) versus female (*n* = 150) participants. Consistent with random assignment, half (*n* = 75) of the experimental and control groups was simulated as half males and half females. We simulated the data to show one item having a higher intercept among females, consistent with well-documented gender differences on specific indicators of internalized disorders (e.g. self-esteem, appetite loss, etc.). Although gender differences in mean levels of depression and anxiety (favoring females) are well documented, here we simulated data consistent with a clinical trial, where all participants were selected to be fully comparable at baseline. For this reason, we simulated the data to show no gender differences in mean levels of

Depression or Anxiety at Pre-Test. We also simulated the data to show differential response to treatment, showing that the previously described reduction in Depression was common to males and females. However, females also showed a significant decrease in anxiety at Post- Test, whereas males showed an increase in anxiety levels at Post-Test, showing a deleterious effect of treatment.

Because we consider the 13-model taxonomy of invariance tests to be an important contribution of ESEM, we have developed an automated, freely available module (https://raw.github.com/pdparker/ESEM/master/exampleScript_ESEMS.R) that allows applied researchers to easily test all 13 models with Mplus through the freeware 'R' software package. For routine applications, applied users need only to use a few lines of code to automatically prepare 13 Mplus files, and to produce a pop-up window with a formatted table with fit statistics for all 13 models. In the worked data example available at https://github.com/pdparker/ESEM we briefly demonstrate its use with simulated data.

Summary

Although CFA seems to have largely superseded EFA, the ICM-CFA measurement model is overly restrictive for most psychological studies, and this problem reinforces applied researcher to use dubious strategies to achieve acceptable fit. For this reason, we recommend that applied researchers routinely compare ESEM and CFA solutions, only retaining CFA if there is good evidence that it is able to fit the data as well as the corresponding ESEM and provides similar estimates of the factor correlations. Because the CFA model is merely a special case of the more general ESEM model, and is nested under it, the comparison of the two solutions is straightforward.

The multigroup and longitudinal approaches to invariance provide a very general and elegant framework for tests of measurement invariance and latent mean differences where 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 multigroup approach is one of the most important applications of ESEM. Essentially, this same approach can be used to evaluate the full 13-model taxonomy of invariance over occasions in longitudinal data. In the related article to these supplementary materials we have reviewed studies that compared factorial solutions generated from ESEM, relative to ICM-CFA. Guided by the taxonomy of invariance models, these studies also demonstrate the utility of ESEM for tests of multiple-group, longitudinal and MIMIC-based measurement invariance. Importantly, our 13-model taxonomy of invariance 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, 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. However, ESEM should generally be preferred to ICM-CFA when the factors are appropriately identified by ESEM, the goodness of fit is meaningful better than for ICM-CFA, and factor correlations are meaningfully smaller than for ICM-CFA. Furthermore, based on Marsh's (2007; Marsh, Hau & Wen., 2004) 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 multigroup and MIMIC tests of measurement invariance, 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 note that autoregressive cross-lagged models and change-score based longitudinal mediation models could be estimated in

ESEM, or within the complementary EwC method.

In that article we also review the published studies that show how ESEM is an overarching integration of the best aspects of CFA/SEM and traditional EFA, with broad applicability to clinical studies not appropriately addressed by either traditional EFAs or CFA/SEMs. However, ESEM is not a panacea, and it may not be appropriate in some situations. Nonetheless, it provides clinical and social science researchers more generally with considerable flexibility to address substantively important issues such as those raised here, where the traditional ICM-CFA approach is not appropriate. Because ESEM is still a relatively new statistical tool, “best practice” will evolve with experience. Nevertheless, the results of this review provide strong support for the application of ESEM in psychological research more generally. Based on our review we recommend that the psychometric evaluation of psychological assessment instruments should routinely apply ESEM and juxtapose the results with the corresponding CFA models that are traditionally used.

Literature Cited

Asparouhov T, Muthén B. 2009. Exploratory structural equation modeling. *Struct. Equ. Modeling.* 16: 397–438

Bandalos DL. 2008. Is parceling really necessary? A comparison of results from item parceling and categorical variable methodology. *Struct. Equ. Modeling*. 15: 211–40

Baron RM, Kenny DA. 1986. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *J. Pers. Soc. Psychol*. 51(6): 1173

Bollen KA. 1989. *Structural Equations with Latent Variables*. New York: Wiley.

Bollen KA. 2002. Latent variables in psychology and the social sciences. *Annu. Rev. Psychol.* 53: 605–34 doi:10.1146/annurev.psych.53.100901.135239

Brown TA. 2006. *Confirmatory Factor Analysis for Applied Research*. New York: Guilford.

Browne MW. 2001. An overview of analytic rotation in exploratory factor analysis. *Multivariate. Behav. Res*. 36: 111–50

Burns G L, Walsh JA, Servera M, Lorenzo-Seva U, Cardo E, Rodríguez-Fornells A. 2013.Construct validity of ADHD/ODD rating scales: Recommendations for the evaluation of forthcoming DSM-V ADHD/ODD scales. *J. Abnorm. Child. Psychol.* 41:15–26 doi:10.1007/s10802-012-9660-5.

Byrne BM. 2011. *Structural Equation Modeling with Mplus: Basic Concepts, Applications, and Programming*. Mahwah, NJ: Routledge.

Byrne BM, Shavelson RJ, Muthén B. 1989. Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychol. Bull*. 105:456–66

Campbell DT, Fiske DW. 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol. Bull*. 56: 81–105

Campbell DT, O'Connell EJ 1967. Methods factors in multitrait-multimethod matrices: Multiplicative rather than additive? *Multivariate. Behav. Res*. 2: 409–26 doi:10.1207/s15327906mbr0204_1

Carroll JB. 1953. An analytical solution for approximating simple structure in factor analysis. *Psychometrika*. 18(1):23-38

Caspi A, Roberts BW, Shiner, RL. 2005. Personality development: Stability and change. *Annu.*

Rev. Psychol. 56:453–84

Cattell, RB. 1949. A note on factor invariance and the identification of factors. *Br. J. Psychol*. 2:134–139

Chen FF. 2007. Sensitivity of goodness of fit indexes to lack of measurement invariance. *Struct. Equ. Modeling*. 14: 464–504

Church AT, Burke PJ. 1994. Exploratory and Confirmatory Tests of the Big 5 and Tellegens 3-Dimensional and 4-Dimensional Models. *J. Pers. Soc. Psychol.*. 66: 93–114

Cohen J. 1968. Multiple regression as a general data-analytic system. *Psychol. Bull.*. 70:426–43

Cheung GW, Rensvold, RB. 2001. The effects of model parsimony and sampling error on the fit of structural equation models. *Organ. Res. Methods*. 4: 236–64

Comrey AL. 1984. Comparison of two methods to identify major personality factors. *Appl. Psychol. Meas*. 8: 397–408

Costa PT, Jr., McCrae RR. 1994 "Set like plaster"? Evidence for the stability of adult personality. In *Can personality change?*, ed. T Heatherton, J Weinberger, pp. 21–40. Washington, DC: American Psychological Association

Cudeck R, MacCallum RC. eds. 2007. *Factor Analysis at 100: Historical Developments and Future Directions*. Mahwah, NJ: Erlbaum.

De Winter JCF, Dodou D, Wieringa PA. 2009. Exploratory factor analysis with small sample sizes. *Multivariate. Behav. Res*. 44: 147–81

Dolan CV, Oort FJ, Stoel, RD, Wichterts, JM. 2009. Testing measurement invariance in the target rotated multigroup exploratory factor model. *Struct. Equ. Modeling*. 16: 295–314

Furnham A, Guenole N, Levine SZ, Chamorro-Premuzic T. 2013. The NEO Personality Inventory–Revised Factor Structure and Gender Invariance From Exploratory Structural Equation Modeling Analyses in a High-Stakes Setting. *Assessment*. 20(1): 14–23

Gallucci M, Perugini M. 2007. The marker index: A new method of selection of marker variables in factor analysis. *TPM-Testing, Psychometrics, Methodology in Applied Psychology*. 14: 3–25

Howarth E. 1972. A factor analysis of selected markers for objective personality factors. *Multivariate. Behav. Res*. 7: 451–76

Hu L, Bentler, PM. 1999. Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Struct. Equ. Modeling*. 6: 1–55

Jöreskog KG. 1969. A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika.* 34: 183–202

Joreskog KG. 1971. Simultaneous factor analysis in several populations. *Psychometrika.* 36: 409–26 doi:10.1007/BF02291366

Jöreskog KG. 1979. Statistical estimation of structural models in longitudinal investigations. In *Longitudinal Research in the Study of Behavior and Development,* ed. JR Nesselroade , B Baltes, pp. 303–51. New York: Academic Press

Jöreskog KG, Goldberger AS. 1975. Estimation of a model with multiple indicators and multiple causes of a single latent variable. *J. Am. Stat. Assoc.* 10: 631–39

Jöreskog KG, Sörbom D. 1979. *Advances in Factor Analysis and Structural Equation Models.* New York: University Press of America

Jöreskog K, Sörbom D. 1993. *LISREL 8: Structural Equation Modeling with the SIMPLIS Command Language.* Chicago, IL: Scientific Software International Inc.

Kushner MG, Maurer EW, Thuras P, Donahue C, Frye B, et al. 2013. Hybrid cognitive behavioral therapy versus relaxation training for co-occurring anxiety and alcohol disorder: A randomized clinical trial. *J. Consult. Clin. Psychol.* 81: 429–42 doi:10.1037/a0031301.

Lang FR, John D, Lüdtke O, Schupp J, Wagner GG. 2011. Short assessment of the Big Five: Robust across survey methods except telephone interviewing. *Behav. Res. Methods.* 43: 548–67 doi:10.3758/s13428-011-0066-z.

Little TD, Card NA, Bovaird JA, Preacher KJ, Crandall CS. 2007. Structural equation modeling of mediation and moderation with contextual factors. In *Modeling contextual effects in longitudinal studies,* ed. TD Little, JA Bovaird, NA Card, pp. 207–30. Mahwah, NJ: Lawrence Erlbaum Associates

Little TD, Cunningham, WA, Shahar, G, Widaman, KF. 2002. To parcel or not to parcel: Exploring the question and weighing the merits. *Struct. Equ. Modeling.* 9: 151–73

MacCallum RC, Zhang S, Preacher KJ, Rucker DD. 2002. On the practice of dichotomization of quantitative variables. *Psychol. Methods.* 7: 19–40

Maïano C, Morin AJS, Lafranchi MC, Therme, P. (In press, Accepted 13 January 2013). The Eating Attitudes Test-26 Revisited using Exploratory Structural Equation Modeling. *J Abnorm Child Psychol.* Online first, doi: 10.1007/s10802-013-9718-z. (ISI 2011: 3.088).

Marsh HW. 1988. Multitrait multimethod analysis. In *Educational Research Methodology, Measurement and Evaluation: An International Handbook*, ed. JP Keeves, pp. 570–80. Oxford: Pergamon Press

Marsh HW. 1993. Multitrait-multimethod analyses: Inferring each trait/method combination with multiple indicators. *Appl. Meas. Educ.* 6: 49–81

Marsh HW. 1994. Confirmatory factor analysis models of factorial invariance: A multifaceted approach. *Struct. Equ. Modeling.* 1: 5–34

Marsh HW. 1995. The analysis of multitrait multimethod data. In *International encyclopedia of education* (2nd ed.), ed. TH Husen, TN Postlethwaite. Oxford: Pergamon Press

Marsh, HW 2007. Application of confirmatory factor analysis and structural equation modeling in sport/exercise psychology. In *Handbook of Sport Psychology* (3rd ed.), ed. G Tenenbaum, RC Eklund, pp. 774–98. New York: Wiley

Marsh HW, Balla JR, Hau KT. 1996. An evaluation of incremental fit indexes: A clarification of mathematical and empirical processes. In *Advanced Structural Equation Modeling Techniques*, ed. GA Marcoulides, RE Schumacker, \pp. 315–53. Hillsdale, NJ: Erlbaum

Marsh HW, Balla JR, McDonald RP. 1988. Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychol. Bull.* 103: 391–410

Marsh, HW Craven, RG 2006. Reciprocal Effects of Self-concept and Performance from a Multidimensional Perspective: Beyond Seductive Pleasure and Unidimensional Perspectives. *Perspect. Psychol .Sci.* 133–63

Marsh HW, Ellis L, Parada L, Richards G, Heubeck BG. 2005. A short version of the Self Description Questionnaire II: Operationalizing criteria for short-form evaluation with new applications of confirmatory factor analyses. *Psychol. Assess.* 17: 81–102

Marsh, HW, Grayson, D. 1994. Longitudinal stability of latent means and individual differences: A unified approach. *Struct. Equ. Modeling.* 1*:* 317–59

Marsh HW, Grayson D. 1995. Latent variable models of multitrait-multi method data. *In Structural Equation modeling: Concepts, Issues, and Applications*, ed. R. H. Hoyle pp. 177–98. Thousand Oaks, CA: Sage Publications

Marsh HW, Hau K-T. 1996. Assessing goodness of fit: Is parsimony always desirable? *J. Exp. Educ.* 64: 364–90

Marsh HW, Hau K-T, Grayson D. 2005. Goodness of Fit Evaluation in Structural Equation

Modeling. In *Psychometrics: A Festschrift to Roderick P. McDonald*, ed. A Maydeu-Olivares, J McCardle. Hillsdale, NJ: Erlbaum

Marsh HW, Hau K-T, Balla JR, Grayson D. 1998. Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate. Behav. Res.* 33:181–220

Marsh HW, Hau KT, 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. *Struct. Equ. Modeling.* 11: 320–41

Marsh HW, Hocevar D. 1988. A new, more powerful approach to multitrait-multimethod analyses: Application of 2nd-order confirmatory factor analysis. *J. Appl. Psychol.* 73: 107–11

Marsh HW, 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 HW, Liem GAD, Martin AJ, Morin AJS, Nagengast B. 2011. Methodological- measurement fruitfulness of exploratory structural equation modeling (ESEM): New approaches to key substantive issues in motivation and engagement. *J. Psychoeduc. Assess.* 29: 322–46

Marsh HW, Lüdtke O, Muthén BO, Asparouhov T, Morin AJS, Trautwein U. 2010. A new look at the big-five factor structure through Exploratory Structural Equation Modeling. *Psychol. Assess.* 22: 471–91

Marsh HW, Lüdtke O, Nagengast B, Morin AJS, Von Davier M. 2013. Why Item Parcels Are (almost) Never Appropriate: Two Wrongs Do Not Make A Right—Camouflaging Misspecification with Item Parcels in CFA Models. *Psychol. Methods*.

Marsh HW, Muthén B, Asparouhov T, Lüdtke O, Robitzsch A, et al. 2009. Exploratory structural equation modeling, integrating CFA and EFA: Application to students' evaluations of university teaching. *Struct. Equ. Modeling.* 16: 439–76

Marsh HW, Nagengast B, Morin AJS. 2013. Measurement invariance of big-five factors over the life span: ESEM tests of gender, age, plasticity, maturity, and la dolce vita effects. *Dev. Psychol.* 49: 1194–218 doi:10.1037/a0026913.

Marsh HW, Nagengast B, Morin AJS, Parada RH, Craven, RG, Hamilton, LR. 2011. Construct validity of the multidimensional structure of bullying and victimization: An application of exploratory structural equation modeling. *J. Educ. Psychol.*. 103: 701–32

Marsh HW, O'Mara A. 2010. Long-term total negative effects of school-average ability on diverse educational outcomes: Direct and indirect effects of the big-fish-little-pond effect. *German J. Educ. Psychol.* 24: 51–72

Marsh HW, O'Neill R. 1984. Self description questionnaire III: the construct validity of multidimensional self- concept ratings by late adolescents. *J. Educ. Meas.* 21(2): 153–74

Marsh HW, Tracey DK, Craven RG. 2006. Multidimensional self-concept structure for preadolescents with mild intellectual disabilities: A hybrid multigroup-MIMIC approach to factorial invariance and latent mean differences. *Educ. Psychol. Meas.* 66: 795–818

Marsh HW, Vallerand RJ, Lafrenière M-AK., Parker P, Morin, AJS, et al. 2013. Passion: Does One Scale Fit All? Construct Validity of Two-Factor Passion Scale and Psychometric Invariance Over Different Activities and Languages. *Psychol. Assess.* Advance online publication. doi: 10.1037/a0032573

Marsh HW, Wen Z, Hau KT, Nagengast 2013. Structural equation models of latent interaction and quadratic effects. In *Structural Equation Modeling: A Second Course* (2nd ed.) ed. G. Hancock, R. Mueller, pp. 267–308. New York: Information Age Publishing

McCrae RR, Costa PT Jr. 1997. Personality trait structure as a human universal. *Am. Psychol.* 52: 509–16

McCrae RR, Zonderman AB, Costa PT Jr, Bond MH, Paunonen S. 1996. Evaluating the replicability of factors in the revised NEO Personality Inventory: Confirmatory factor analysis versus Procrustes rotation. *J. Pers. Soc. Psychol.* 70: 552–66

McDonald RP. 1985. *Factor analysis and related methods.* Hillsdale, NJ: Erlbaum.

Meleddu M, Guicciardi M, Scalas LF, Fadda D. 2012. Validation of an Italian version of the Oxford Happiness Inventory in adolescence. *J. Pers. Assess.* 94: 175–85

Meredith W. 1964. Rotation to achieve factorial invariance. *Psychometrika.* 29: 187–206. Meredith W. 1993. Measurement invariance, factor analysis and factorial invariance. *Psychometrika* 58: 525–43

Millsap RE. 2011. *Statistical Approaches to Measurement Invariance.* New York: Routledge

Mora PA, Berkowitz A, Contrada RJ, Wisnivesky J, Horne R, Leventhal H, Halm EA. 2011. Factor structure and longitudinal invariance of the Medical Adherence Report Scale-Asthma. *Psychol. Health.* 26(6): 713–27

Morin AJS, Maïano C. 2011. Cross-validation of the Short Form of the Physical Self-Inventory (PSI-S) using Exploratory Structural Equation Modeling (ESEM). *Psychol. Sport. Exerc*. 12: 540–54

Morin AJS, Marsh HW, Nagengast B. 2013. Exploratory structural equation models. In *Handbook of Structural Equation Modeling*, ed. R Hoyle, S West. Guilford Press, NY Muthén BO. 1989. Latent variable modeling in heterogenous populations. *Psychometrika* 54:557–85

Muthén BO. 2002. Beyond SEM: General latent variable modeling. *Behaviormetrika* 29: 81–117

Muthén B, Asparouhov T. 2012. Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychol. Methods.* 17: 313–35 doi:10.1037/a0026802

Muthén LK, Muthén BO. 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 LK, Muthén B. 2011. *Mplus User's Guide*. Los Angeles CA: Muthén & Muthén

Myers ND, Chase MA, Pierce SW, Martin E. 2011. Coaching efficacy and exploratory structural equation modeling: A substantive-methodological synergy. *J. Sport. Exerc. Psychol*. 33: 779–806

Nagengast B, Marsh HW. 2012. Big Fish in Little Ponds Aspire More: Mediation and Cross-Cultural Generalizability of School-Average Ability Effects on Self-Concept and Career Aspirations in Science. *J. Educ. Psych.* Advance online publication. doi: 10.1037/a0027697

OECD (Organisation for Economic Co-operation and Development). (2007). PISA 2006 science competencies for tomorrow's world. Paris: Author.

Olson JF, Martin MO, Mullis, IVS. eds. 2008. *TIMSS 2007 Technical Report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center.

Overall JE. 1974. Marker variable factor analysis: A regional principal axes solution. *Multivariate. Behav. Res.* 9: 149–64.

Pettersson E, Turkheimer E, Horn E, Menatti AR. 2012. The general factor of personality and evaluation. *Eur. J. Pers.* 26: 292–302 doi:10.1002/per.839.

Reise SP. 2012. The rediscovery of bifactor measurement models. *Multivariate. Behav. Res.* 47: 667–96

Sass DA, Smith PL. 2006. The effects of parceling unidimensional scales on structural parameter estimates in structural equation modeling. *Struct. Equ. Modeling*. 13:566–86

Sass DA, Schmitt TA. 2010. A comparative investigation of rotation criteria within exploratory factor analysis. *Multivariate. Behav. Res.* 45: 1–33

Schmitt TA, Sass DA. 2011. Rotation criteria and hypothesis testing for exploratory factor analysis: Implications for factor pattern loadings and interfactor correlations. *Educ. Psychol. Meas.* 71: 95–113

Skrondal A, Rabe-Hesketh S. 2004. *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. New York: Chapman & Hall/CRC.

Strauss ME, Smith GT. 2009. Construct validity: Advances in theory and methodology. *Annu. Rev. Clin. Psychol.* 5: 1–25 doi:10.1146/annurev.clinpsy.032408.153639.

Thurstone LL. 1947. *Multiple Factor Analysis*. Chicago: University of Chicago.

Toland MD, De Ayala RJ. 2005. A Multilevel factor analysis of students' evaluations of teaching. *Educ. Psychol. Meas.* 65: 272–96

Tomarken AJ, Waller NG. 2005. Structural equation modeling: Strengths, limitations, and misconceptions. *Annu. Rev. Clin. Psychol.*. 1: 31–65

Vandenberg RJ, Lance CE. 2000. A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organ. Res. 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? *Eur. J. Pers.* 11: 147–66.

Velicer WF, Fava JL. 1998. The effects of variable and subject sampling on factor pattern recovery. *Psychol. Methods*. 3: 231–51

Williams LJ, O'Boyle EH Jr. 2008. Measurement models for linking latent variables and indicators: A review of human resource management research using parcels. *Human Resource Manag. Rev.* 18: 233–24