

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

Herbert W. Marsh,<sup>1,2,3</sup> Alexandre J.S. Morin,<sup>1</sup>  
Philip D. Parker,<sup>1</sup> and Gurvinder Kaur<sup>1</sup>

<sup>1</sup>Department of Education, University of Western Sydney, Penrith NSW 2751, Australia;  
email: h.marsh@uws.edu.au

<sup>2</sup>Department of Education, University of Oxford, Oxford, United Kingdom OX2 6PY

<sup>3</sup>King Saud University, School of Education, Riyadh, Saudi Arabia 11451

Annu. Rev. Clin. Psychol. 2014. 10:85–110

First published online as a Review in Advance on  
December 2, 2013

The *Annual Review of Clinical Psychology* is online at  
clinpsy.annualreviews.org

This article's doi:  
10.1146/annurev-clinpsy-032813-153700

Copyright © 2014 by Annual Reviews.  
All rights reserved

## Keywords

exploratory and confirmatory factor analysis, exploratory structural equation models, exploratory structural equation model within confirmatory factor analysis, multiple-indicator multiple-cause (MIMIC) models, multitrait-multimethod models, bifactor models

## Abstract

Exploratory factor analysis (EFA) and confirmatory factor analysis (CFA), path analysis, and structural equation modeling (SEM) have long histories in clinical research. Although CFA has largely superseded EFA, CFAs of multidimensional constructs typically fail to meet standards of good measurement: goodness of fit, measurement invariance, lack of differential item functioning, and well-differentiated factors in support of discriminant validity. Part of the problem is undue reliance on overly restrictive CFAs in which each item loads on only one factor. Exploratory SEM (ESEM), an overarching integration of the best aspects of CFA/SEM and traditional EFA, provides confirmatory tests of a priori factor structures, relations between latent factors and multigroup/multioccasion tests of full (mean structure) measurement invariance. It incorporates all combinations of CFA factors, ESEM factors, covariates, grouping/multiple-indicator multiple-cause (MIMIC) variables, latent growth, and complex structures that typically have required CFA/SEM. ESEM has broad applicability to clinical studies that are not appropriately addressed either by traditional EFA or CFA/SEM.

## Contents

EXPLORATORY STRUCTURAL EQUATION MODELING AS AN INTEGRATIVE FRAMEWORK: AN INTRODUCTION .....	86
THE EXPLORATORY STRUCTURAL EQUATION MODELING APPROACH .....	89
Identification and Rotational Indeterminacy .....	89
The Size of Factor Correlations and Discriminant Validity .....	91
Extending ESEM: The ESEM-Within-CFA Approach .....	92
Measurement Invariance .....	92
OVERVIEW OF PUBLISHED EXPLORATORY STRUCTURAL EQUATION MODELING APPLICATIONS WITH RELEVANCE TO CLINICAL AND SOCIAL SCIENCE RESEARCH .....	95
Content Analysis of Published ESEM Applications .....	95
Illustrative Applications Demonstrating Initial or Novel Applications of ESEM ....	96
DIRECTIONS FOR FUTURE DEVELOPMENT .....	103
Using ESEM Factors in Subsequent Analyses: Manifest Scores and Factor Scores Versus Latent Correlation Matrices and/or Plausible Values .....	103
Juxtaposing EFA, CFA, ESEM, and Bayesian Structural Equation Models? .....	104
Recommendations for Applied Clinical Researchers .....	104

## EXPLORATORY STRUCTURAL EQUATION MODELING AS AN INTEGRATIVE FRAMEWORK: AN INTRODUCTION

Historically, researchers have relied on exploratory factor analyses (EFAs) to identify and distinguish between key psychological constructs, but many analyses that are central to clinical research cannot be easily performed with EFA. For example, in EFA it is not easy to test measurement invariance (in relation to groups, time, and covariates), which is the assumption in many research designs, such as randomized control trials, or to incorporate latent EFA factors into subsequent analyses, relating them to other constructs, to interventions, or to changes over time. Hence, clinical researchers typically have to resort to suboptimal, nonlatent (manifest) scale or factor score representations of latent EFA factors, followed by using manifest statistical models [e.g., t-tests, analyses of variance (ANOVAs), or multiple regressions] to test for relationships between these manifest scores and other variables or interventions. Cohen's (1968) seminal publication presented multiple regression as a sufficiently general framework to incorporate traditional univariate and multivariate analyses of manifest variables [e.g., t-tests, ANOVAs, multivariate analyses of variance (MANOVAs)] as special cases of multiple regression. Although highly flexible, this framework still could not incorporate latent variables corrected for measurement errors, so latent psychometric constructs identified through EFA still had to be converted to suboptimal scale or factor scores. The advent of confirmatory factor analysis (CFA)/structural equation modeling (SEM) made it possible to conduct systematic tests of measurement invariance (e.g., Jöreskog & Sörbom 1979, Meredith 1993) and led to many additional advances, including the analysis of relationships involving latent constructs estimated after correction for measurement error. The basic independent clusters model of confirmatory factor analysis (ICM-CFA) posits that all items have zero factor loadings on all factors other than the one they are designed to measure (McDonald 1985). Following from seminal work by Jöreskog & Sörbom (1979) and others, researchers (e.g., Muthén 2002,

**EFA:** exploratory factor analysis

**Manifest variable:** a variable directly observed/measured or defined by a single indicator (although this may be an average of multiple indicators)

**Latent variable:** an unobserved hypothetical construct; in factor analysis, typically defined in relation to multiple indicators

**CFA:** confirmatory factor analysis

Skrondal & Rabe-Hesketh 2004) integrated these features into an even more generic framework (generalized SEM), allowing for the estimation of relations between any manifest and latent continuous or categorical variables. Indeed, Tomarken & Waller (2005) have highlighted the importance of CFA/SEM in clinical psychology, noting the large number of publications that indicate it has become the most commonly used multivariate technique. Here we present exploratory structural equation modeling (ESEM) as an even more general framework that incorporates CFA/SEM and EFA as special cases, and we demonstrate why ESEM is typically preferable to the more restricted CFA/SEM in clinical psychology research.

Although EFA is an important precursor of CFA/SEM (Cudeck & MacCallum 2007), it is widely seen as less useful, partly on the basis of the semantically based misconception that it is purely an “exploratory” method that should be used only when the researcher has no a priori assumption regarding factor structure. Thus, for example, in his review of latent variable models for the *Annual Review of Psychology*, Bollen (2002, p. 615) noted that:

In exploratory factor analysis, the factors are extracted from the data without specifying the number and pattern of loadings between the observed variables and the latent factor variables. In contrast, confirmatory factor analysis specifies the number, meaning, associations, and pattern of free parameters in the factor loading matrix before a researcher analyzes the data.

Similarly, in the *Annual Review of Clinical Psychology*, Strauss & Smith (2009, pp. 16–17) noted that:

A major advantage of CFA in construct validity research is the possibility of directly comparing alternative models of relationships among constructs, a critical component of theory testing.

However, such oversimplified distinctions camouflage the critical difference: that all cross-loadings traditionally constrained to be zero in CFA are freely estimated in EFA, so ICM- CFA structures are much more restrictive than EFA structures. Because of this, in many instances item-level CFAs fail to provide clear support for instruments that apparently had been well established in EFA research (e.g., Marsh et al. 2009, 2010). In illustration of this, Marsh (2007; Marsh et al. 2005) proposed an intentionally extreme “straw person” claim that should have been easy to refute with empirical evidence:

It is almost impossible to get an acceptable fit (e.g., CFI, TLI > 0.9; RMSEA < 0.05) 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. (Marsh 2007, p. 785)

However, when Marsh placed this claim on SEMNET (an electronic network devoted to SEM) and invited the more than 2,000 members to provide published counterexamples, no one was able to do so. This suggests that many psychological instruments routinely used in applied research do not even meet the minimum criteria of acceptable fit, based on current ICM-CFA standards.

Marsh and colleagues (2009, 2010, 2011a,b, 2013b; also see Morin et al. 2013) argue that ICM-CFAs are too restrictive. Factor structures based on measures used in applied research typically include cross-loadings that can be justified by substantive theory or by item content (e.g., method effects), or that simply represent another source of measurement error, whereby items are fallible indicators of the constructs and thus tend to have small residual associations with other constructs (Asparouhov & Muthén 2009, Church & Burke 1994). In some cases these might be eliminated

---

#### Structural equation model (SEM):

a combination of the measurement (CFA) model of exogenous constructs not influenced by other variables and the structural model of directed (predictive) paths relating latent and/or manifest variables

#### ICM-CFA:

independent clusters model of confirmatory factor analysis

**ESEM:** exploratory structural equation modeling

---

### Correlated

**uniquenesses:** the covariances between residual items' variance terms not explained by the theoretical constructs often associated with their use in longitudinal data or method effects associated with parallel-worded items

### MIMIC:

multiple-indicator multiple-cause

### Measurement model:

CFA model of exogenous constructs not influenced by other variables in the model with no causal (or predictive) paths. Structural models with causal paths are typically equivalent to or nested under measurement models

**Supplemental Material**

in part by the development of psychometrically stronger measures, but it is our contention that most items have multiple determinants, so nonzero cross-loadings are inherent in psychological measurement and can often be logically anticipated from the nature of the items themselves (for instance, many clinical symptoms of psychological disorders can be associated with multiple diagnostic categories: either as symptoms or as associated characteristics). These small cross-loadings are important because requiring them to be zero typically results in inflated CFA factor correlations that detract from the discriminant validity of the factors and lead to biased estimates in SEMs incorporating other variables (Asparouhov & Muthén 2009; Marsh et al. 2009, 2010; Schmitt & Sass 2011). Furthermore, the strategies often used to compensate for these problems in CFA (e.g., parceling, ex post facto modifications such as ad hoc correlated uniquenesses) tend to be counterproductive, 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? The answer, apparently, is the mistaken belief that many recent advances in latent variable modeling require CFA/SEMs. Here we outline ESEM (Asparouhov & Muthén 2009; Marsh et al. 2009, 2010; Morin et al. 2013), an integration of EFA, CFA, and SEM that has the potential to resolve this dilemma and has wide applicability to clinical research. We assume that readers are reasonably familiar with EFA, CFA, and SEM (otherwise, for an introduction see Bollen 1989, Brown 2006, Byrne 2011, Cudeck & MacCallum 2007).

ESEM shares many characteristics with CFA that fundamentally distinguish it from traditional approaches to EFA, such as tests of predictive relations between latent constructs adjusted for measurement error, method factors, correlated uniquenesses, complex error structures, bifactor models, full measurement invariance over groups or occasions, latent mean structures, differential item functioning (i.e., noninvariance of item intercepts), extension of factor analysis to SEMs, auto-regressive path models of causal ordering, and multiple-indicator multiple-cause (MIMIC) models of relations of latent factors with background and predictor variables. Owing to space limitations, and because all of these features normally associated with CFA/SEM are covered elsewhere in detail, here we touch on them only briefly as they relate to ESEM (for additional information, follow the **Supplemental Material link** in the online version of this article or at <http://www.annualreviews.org/>). Rather, we emphasize the important limitations of traditional ICM-CFA models in many applied studies, which are overcome by using ESEM. These limitations include poor fit to item-level factor structures, poor discriminant validity associated with inflated correlations among CFA factors, and biased structural parameter estimates in SEMs based on misspecified measurement models. We then provide an overview of published ESEM studies, illustrate some new developments, and conclude with a discussion of limitations and directions for further studies.

It is also important to note that EFAs and CFA/SEMs are only special cases of more general ESEMs (e.g., Morin et al. 2013). In particular, EFA factors are ESEM factors. Although EFAs are often seen as exploratory, we view ESEM as a primarily confirmatory approach, and the use of target rotation formalizes this view, as it allows the analyst much more a priori control on the expected factor structure. However, traditional EFA and ESEM can both be used as exploratory or confirmatory tools (as, indeed, can CFA/SEM), depending on the nature of the research application, theory, and data.

Because of space limitations, we are not able to include the details of worked examples (i.e., data, syntax, and discussion of results), but we have created a separate website with an expanded discussion of a data set simulated to reflect a typical clinical application (<https://github.com/pdparker/ESEM>). This data set includes six items serving as indicators of two correlated factors (anxiety and depression), two correlated/comorbid clinical states that can be measured by indicators/symptoms that realistically can be expected to present significant cross-loadings. Simulating a clinical pretest

posttest design with randomized experimental and control groups, we also simulated a second set of six parallel posttest items where the factor structure differed slightly from the pretest data. The control group was simulated to show a small decrease of depressive symptoms over time, but no change in anxiety levels, whereas the experimental group showed a substantial decrease in depressive symptoms over time, with gender-differentiated effects regarding the response to treatment for symptoms of anxiety (i.e., construct-specific intervention effects). Readers are invited to explore these examples for a more detailed understanding of ESEM and its relevance to clinical research.

---

**Nested:** models that are obtained by placing restrictions on another model; parameters in one can be represented as a subset of those in the other

---

## THE EXPLORATORY STRUCTURAL EQUATION MODELING APPROACH

### Identification and Rotational Indeterminacy

**Identification.** All parameters in ESEM can be identified with the maximum likelihood (ML) estimator, with weighted least square estimators, or with robust alternatives. In ESEM, multiple sets of ESEM factors can be defined either as ESEM or CFA factors. 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 is usually determined on the basis of a priori theoretical expectations, on practical considerations, or perhaps posthoc, based on preliminary tests conducted on the data. The integrative framework provided by ESEM is demonstrated, in that ESEM is appropriate for any combination of ESEM and CFA factors and is easily extended to accommodate predictive SEMs involving ESEM and CFA factors.

If the ESEM model includes a single factor or only ICM-CFA factors, then it is equivalent to the classic CFA/SEM model. When the general ESEM model contains more than one ESEM factor ( $m > 1$ ) with cross-loadings, a different set of constraints is required to achieve an identified solution (for further discussion, see Asparouhov & Muthén 2009; Marsh et al. 2009; 2010; Sass & Schmitt 2010). In the first step, an unconstrained factor structure is estimated in which a total of  $m^2$  constraints is required to achieve identification (Jöreskog 1969). 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). Because the basic ICM-CFA model is nested under the corresponding ESEM, conventional approaches to model comparison can be used to compare the fit of the two models—along with a detailed evaluation of parameter estimates based on the two approaches. ESEM is most appropriate when it fits the data better than does a corresponding CFA model. Otherwise, the CFA factor structure is preferable, on the basis of parsimony (Marsh et al. 2013b). However, a growing body of research suggests that ICM-CFA models are typically too restrictive to provide an acceptable fit for many psychological instruments (Marsh 2007, Morin et al. 2013).

Early applications of ESEM (Marsh et al. 2009, 2010) were based on a geomin rotation that was developed to represent Thurstone's (1947) simple structure and to incorporate a complexity parameter ( $\epsilon$ ) that increases with the number of factors (Asparouhov & Muthén 2009, Browne 2001). Although Marsh et al. (2009, 2010) used an  $\epsilon$  value of 0.5 with complex measurement instruments so as to avoid inflated factor correlations, Asparouhov & Muthén (2009) recommend comparing solutions based on varying  $\epsilon$  values. More recently, Marsh, Lüdtke, and colleagues (2010, Marsh et al. 2013a) recommended the use of target rotation, particularly when a few items from each factor are relatively pure measures of the factor (i.e., factor cross-loadings are near zero). As emphasized by Browne (2001; also see Asparouhov & Muthén 2009, Dolan et al. 2009),



---

**Goodness of fit:** indices that evaluate how well a posited model fits the data: e.g., the Comparative Fit Index (CFI), the Tucker-Lewis Index (TLI), and the root mean square error of approximation (RMSEA)

---

this strategy reflects a compromise between the mechanical approach to EFA rotation and the a priori ICM-CFA restrictive model, based on partial knowledge of the factor structure, and is consistent with the view that ESEM is more typically used for confirmatory rather than exploratory purposes. The target rotation is particularly appropriate when there is a clearly defined a priori factor structure and a reasonable approximation to simple structure.

The identification strategy for the ESEM mean structure is similar to typical CFAs: Item intercepts are freely estimated and latent factor means are constrained to zero (due to rotational difficulties, the alternative CFA method of constraining one intercept per factor to zero to freely estimate the latent means is not recommended in ESEM). In the standard ESEM model, all of these constraints are the default in the estimation process where, in addition, multiple random starting values are employed to help protect against nonconvergence and local minima. For a detailed presentation of identification and estimation issues, readers are referred to Asparouhov & Muthén (2009), Marsh et al. (2009, 2010), and Sass & Schmitt (2010).

**Rotational indeterminacy.** A potentially important limitation exists with ESEMs and EFAs in that the pattern of cross-loadings and the size of the estimated factor correlations vary with the specific rotation (e.g., Browne 2001, Sass & Schmitt 2010). Because rotation is independent of goodness of fit, and different rotations all fit the data equally well, goodness of fit provides no basis for choosing the best rotation (Sass & Schmitt 2010, Schmitt & Sass 2011). On the basis of simulated data, Marsh et al. (2013a) argued that this issue is circumvented to some extent by target rotation. However, this was based on population-generating models in which some of the items representing each factor had zero cross-loadings in the population-generating model and were target items in the target rotation. Target rotation does not require there to be anchor items with zero nontarget factor loadings, but having them—or at least a reasonable approximation of simple structure—provides a stronger a priori model, gives the researcher greater control in specifying the model, and facilitates interpretation of the results. Although it is common in factor analysis for several of the indicators to serve as “markers” of the factor (e.g., Cattell 1949, Comrey 1984, Gallucci & Perugini 2007, Howarth 1972, Overall 1974), this is clearly not the case in all situations, and the target rotation might not be expected to perform as well in all cases, at least in terms of accurately estimating the true population correlation.

The historical rationale for most rotation strategies has been based on maximizing the simple structure of the factor loadings (either for variables, factors, or a combination of the two) with little regard to the appropriateness of factor correlation estimates (Sass & Schmitt 2010). 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 simple structures in which nontarget loadings are ideally small relative to target loadings but are not required to be zero (Carroll 1953, McDonald 1985, Thurstone 1947). Indeed, as emphasized by Sass and Schmitt, there is necessarily a balance between constraints on the sizes of cross-loadings and factor correlations (one extreme being the ICM-CFA solution, which constrains cross-loadings to be zero and typically results in substantially inflated factor correlations when this assumption is violated and the other extreme, orthogonal rotation, in which all correlations are constrained to be zero and which typically results in substantially inflated cross-loadings). Morin & Maïano (2011) systematically illustrate the issue of rotational indeterminacy in relation to how factor correlations are modified as a function of the rotation procedure. Although resolution of this problem of choosing the most appropriate rotation strategy is clearly beyond the scope of this review, it is important to emphasize that the goodness of fit for the ESEM solution does not depend on the rotation. Hence, if the fit of the ESEM solution is substantially better than that of the ICM-CFA solution, then the estimated correlation for the ICM-CFA solution is likely to be substantially biased. Nevertheless, pending further research,

we recommend that researchers compare the results based on alternative rotational procedures or provide clear arguments for their choice of rotational method.

## The Size of Factor Correlations and Discriminant Validity

Marsh and colleagues (2010, 2013a) argued that the ICM-CFA factor correlations are likely to be positively biased—sometimes substantially—unless nontarget loadings are close to zero, as consistently shown in simulation (e.g., Asparouhov & Muthén 2009, Marsh et al. 2013a) and real data studies (e.g., Marsh et al. 2011b). In relation to clinical and psychological research, this positive bias 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 diagnostic usefulness that depends on having well-differentiated factors. Furthermore, this bias estimation of factor correlations affects results in other parts of SEMs that are not easy to predict a priori. This has been a potentially serious problem in applied research, where the primary focus is on relations among the factors and their relations with other constructs (e.g., background variables, covariates, interventions, or subsequent outcomes). 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 overparameterization of the model, is unlikely to result in bias in factor correlations.

Using a combination of real and simulated data, Marsh et al. (2013a) provide perhaps the strongest evidence of the bias in factor correlation estimates based on ICM-CFA factors. Based on responses to 24 items [neuroticism and extraversion from the Neuroticism-Extraversion-Openness Five-Factor Inventory (NEO-FFI) Big Five personality instrument] using a target rotation, the ESEMs fit substantially better than CFAs. Of particular relevance, the factor correlation was substantially smaller for ESEM ( $r = 0.15$ ) than for the corresponding CFA solution ( $r = 0.51$ ) and more consistent with theoretical predictions that the Big Five personality factors are reasonably orthogonal. Noting that the use of real data precluded knowledge of the true population correlation, they then simulated data in which the true population correlation was either 0.25 or 0.60, based on one of four simple-structure solutions: a pure ICM-CFA (all nontarget loadings = 0), nearly pure ICM-CFA (nontarget loadings 0 or 0.1), approximate (cross-loadings 0, 0.1, or 0.2), and moderate (cross-loadings 0–0.4). For the pure ICM-CFA data, both CFAs and ESEMs fitted the data, and both accurately estimated the population correlation, but the CFA was considered the best on the basis of parsimony. However, even for the nearly pure ICM-CFA structure, the CFA that failed to take into account the (very small) cross-loadings resulted in inflated estimates of the known population correlation:  $r = 0.41$  (for  $\rho = 0.25$ ) and  $r = 0.71$  (for  $\rho = 0.60$ ). For the moderate and approximate solutions, the bias was substantially higher:  $r$ 's = 0.52 and 0.84 (for  $\rho = 0.25$ ) and  $r$ 's = 0.78 and 0.94 (for  $\rho = 0.60$ ). In each case, ESEM provided an almost perfect fit that accurately estimated the factor correlation. Thus, Marsh et al. (2013a) argued that both ESEM and ICM-CFAs should routinely be applied to the same data.

In this same article, Marsh et al. (2013a) argue against the widespread practice of using parcels instead of items. Marsh et al. (1998; also see De Winter et al. 2009, Velicer & Fava 1998) argued that “more is never too much” for the number of indicators, as well as the number of participants, as generalizability is typically enhanced by having larger samples of participants and items. Historically, it has been common to have 10 to 15+ items per scale on the most widely used psychological tests, but there is an understandable reluctance to incorporate large numbers of indicators into complex CFA/SEMs. One widely used compromise (e.g., Little et al. 2002, Marsh et al. 1988) is to collect many items but to use item parcels in the analyses. For example, in a psychological instrument

---

### NEO-FFI:

Neuroticism, Extraversion, and Openness Five-Factor Inventory of personality; the 60-item version of the NEO instrument

### Item parcels

(testlets): averaging sets of items from one factor to create a smaller number of indicators (e.g., a twelve-item scale transformed into four three-item parcels)

---

**EwC:** ESEM within CFA

**Mplus:** widely used statistical package that tests ESEM structures

assessing 10 factors with 12 items each, the 120 items could be used to form three four-item parcels for each factor used in the analysis. The critical assumption underlying the appropriate use of parcels is well established (e.g., Bandalos 2008; Little et al. 2002; Marsh & O'Neill 1984; Marsh et al. 1998, 2013a; Sass & Smith 2006; Williams & O'Boyle 2008): Responses to each different factor must be purely unidimensional, with no nonzero cross-loading—in short, an ICM-CFA model fits the data at the item level. However, Marsh et al. (2013a) argued that the use of item parcels is almost never appropriate because (a) the basic assumption of pure unidimensionality is rarely met; (b) biased parameter estimates (e.g., inflated factor correlations) evident in analyses at the item level are not corrected; and (c) results provide such misleadingly good fit indexes that applied researchers, reviewers, and readers might be misled into believing that misspecification problems are resolved. They showed that item parcels are only appropriate if ESEMs and ICM-CFAs both fit the data well and are similar. Although tests of unidimensionality are sometimes given token lip service in justifying the use of parcels, Williams & O'Boyle (2008) emphasize that a primary motivation for their use is the typically unstated need to meet seemingly traditional criteria of acceptable fit even when misfit in analyses at the item level is so great that fits are not acceptable. Of critical importance, Marsh et al. (2013a) showed that the inflated correlations in ICM-CFA factors due to constraining all cross-loadings to be zero were also evident in parcel solutions but not in ESEM solutions based on items.

### Extending ESEM: The ESEM-Within-CFA Approach

The ESEM approach is very flexible, but currently its operationalization still presents some limitations compared to CFA/SEMs (also see Asparouhov & Muthén 2009; Marsh et al. 2009, 2010). For example, all of the factors forming a set of ESEM factors need to be simultaneously related or unrelated to other variables in the model, and tests of the partial invariance of factor loadings are not allowed (though partial invariance of uniquenesses and item intercepts is possible). Marsh et al. (2013b, Morin et al. 2013) proposed a method they called ESEM within CFA (EwC) to circumvent these and related problems.

EwC is an extension of an initial proposal by Jöreskog (1969; also see Muthén & Muthén 2009, slides 133–146) that provides a solution to some of the aforementioned limitations of ESEM. The EwC model must contain the same number of restrictions as the ESEM model (i.e.,  $m^2$  restrictions where  $m$  = number of factors; see previous discussion). In the EwC approach (Marsh et al. 2013b, Morin et al. 2013) all parameter estimates from the final ESEM solution should be used as starting values to estimate the EwC model. A total of  $m^2$  constraints need to be added for this model to be identified. This is most easily accomplished by merely retaining the pattern of fixed and free parameters in the initial ESEM solution, using ESEM estimates for starting and fixed values. (This is greatly facilitated in Mplus v7.1, which allows researchers to copy syntax—including start values for fixed and estimated parameters—as part of ESEM.) The EwC solution is equivalent to the ESEM solutions in that it has the same degrees of freedom, goodness of fit, and parameter estimates as the ESEM solution. Importantly, the researcher has more flexibility in terms of constraining or further modifying the EwC model (because it is a true CFA model) than with the ESEM model upon which it is based (also see **Supplemental Material**; follow the **Supplemental Material link** in the online version of this article or at <http://www.annualreviews.org/>); this provides a useful complement to ESEM and overcomes what were thought to be limitations of ESEM.

### Measurement Invariance

Of particular substantive importance for clinical research are mean-level differences across multiple groups (e.g., male versus female groups, various age groups, clinical versus nonclinical populations;



treatment versus control groups) or over time (i.e., observing the same group of participants on multiple occasions, perhaps before and after an intervention). Tests of whether the underlying factor structure is the same for different groups or occasions have often 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 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.

**Measurement invariance and latent mean comparisons.** 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 is an important prerequisite to any valid form of group-based comparison. Historically, multigroup tests of invariance were seen as a fundamental advantage of CFA/SEM over EFA approaches, which were largely limited to descriptive comparisons of the factor loadings estimated separately in each group (but see Dolan et al. 2009 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 with factor invariance (e.g., Jöreskog & Sörbom 1993; Marsh 1994, 2007; Marsh & Grayson 1994) and item-response-theory approaches to measurement invariance (e.g., Meredith 1964, 1993; also see Millsap 2011, Vandenberg & Lance 2000). Key models test goodness of fit with

**Table 1** Taxonomy of multigroup tests of invariance testable with exploratory structural equation modeling and nesting relations (in brackets)

Model	Parameters constrained to be invariant
Model 1	None (configural invariance)
Model 2	Factor loadings (FL) [1] (weak factorial/measurement invariance)
Model 3	FL uniquenesses (uniq) [1, 2]
Model 4	FL, factor variance-covariances (FVCV) [1, 2]
Model 5	FL, intercepts (inter) [1, 2] (strong factorial/measurement invariance)
Model 6	FL, uniq, FVCV [1, 2, 3, 4]
Model 7	FL, uniq, inter [1, 2, 3, 5] (strict factorial/measurement invariance)
Model 8	FL, FVCV, inter [1, 2, 4, 5]
Model 9	FL, uniq, FVCV, inter [1–8]
Model 10	FL, inter, factor means (FMn) [1, 2, 5] (latent mean invariance)
Model 11	FL, uniq, inter, FMn [1, 2, 3, 5, 7, 10] (manifest mean invariance)
Model 12	FL, FVCV, inter, FMn [1, 2, 4, 5, 6, 8, 10]
Model 13	FL, uniq, FVCV, inter, FMn [1–12] (complete factorial invariance)

Bracketed values represent nesting relations in which the estimated parameters of the less general model are a subset of the parameters estimated in the more general model under which it is nested. All models are nested under Model 1 (with no invariance constraints), whereas Model 13 (complete invariance) is nested under all other models.

no invariance constraints (configural invariance, Model 1); invariance of factor loadings (weak measurement invariance, Model 2) alone or in combination with invariance of factor correlations (factor variance-covariance invariance, Model 4), item intercepts (strong measurement invariance, Model 5), or item intercepts and measurement error (strict measurement invariance, Model 7). 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.

Essentially the same logic and 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 between responses to the same item on different occasions (see Jöreskog 1979, Marsh 2007, Marsh & Hau 1996). Although occasions are the most typical test of invariance over a within-subject construct, this is easily extended to include other within-subject variables (e.g., spouse, therapist, and social worker ratings of the same patient). Indeed, it is possible to extend these models to test the invariance over multiple grouping variables or combinations of multigroup (between-subject) and within-subject variables. Although these tests in each of the 13 models posit full invariance of all parameter estimates for all groups or occasions, Byrne et al. (1989, also see Marsh 2007) have argued for the usefulness of a less demanding test of partial invariance in which a subset of parameters is not constrained to be invariant.

Our 13-model taxonomy is more extensive than most treatments of invariance and was especially designed for ESEM (Marsh et al. 2009), but it is important to emphasize that all these models can be tested with either ESEM or CFA. However, unless the ICM/CFA model is able to fit the data as well as the corresponding ESEM model, ESEM invariance tests offer a viable alternative that overcomes a potentially overly restrictive ICM/CFA structure. Indeed, the ability of ESEM to provide such a rich set of invariance tests of an EFA measurement structure is a remarkable contribution and clearly reinforces the confirmatory nature of ESEM. Because we consider the 13-model taxonomy of invariance tests to be such an important contribution of ESEM, we have developed an automated, freely available module (available at <http://raw.github.com/pdparker/ESEM>) that allows applied researchers to easily test all 13 models with Mplus through the freeware “R” software package.

**The MIMIC approach to prediction, measurement invariance, and differential item functioning.** The multigroup approach to invariance is most appropriate for variables that are naturally categorical (e.g., gender, diagnostic categories, treatment groups) but might not be practical for continuous variables (e.g., age), for studies that evaluate simultaneously many different contrast variables and their interactions, or when sample sizes are small. Although it is always possible to categorize continuous variables into a small number of discrete categories, it is well known in psychological research that this strategy has potentially serious limitations in the reduction of reliability and power (MacCallum et al. 2002), particularly when the continuous predictor variable might have nonlinear effects. The MIMIC model (see **Supplemental Material** for further discussion; follow the **Supplemental Material link** in the online version of this article or at <http://www.annualreviews.org/>) provides an alternative multigroup invariance approach to measurement invariance and differential item functioning by (Morin et al. 2013):

- saturated MIMIC models with paths from each predictor variable and all the item intercept terms, but not the latent factors, and
- invariant intercept MIMIC models with freely estimated paths from the predictor variables to latent factors, but with paths to item intercepts all constrained to be zero.

If the saturated MIMIC model fits substantively better than the intercept-invariant MIMIC model, then there is evidence of differential item functioning (i.e., noninvariance of intercepts).

However, the MIMIC approach is limited in that it assumes the invariance of factor loadings and uniquenesses but does not easily allow for the verification of these assumptions. Hence, both the multiple-group and MIMIC approaches to invariance have contrasting limitations. Thus, Marsh and colleagues (2006) proposed a hybrid approach in which multigroup models (e.g., age groups as discrete categories) are used to test invariance assumptions that cannot easily be tested with the MIMIC approach, and the MIMIC approach (e.g., age as a continuous variable, perhaps representing linear and nonlinear components) is used to infer differences in relation to a score continuum and interactions. Thus, age is treated as a categorical variable with a relatively small number of discrete categories in the multiple-group approach but as a continuous variable in the MIMIC approach. So long as the two approaches converge to similar interpretations, there is support for the construct validity of interpretations based on either approach. Within the context of ESEM, this hybrid approach has been extended to incorporate both the MIMIC and the multiple-group approaches into a single model (see subsequent discussion in Marsh et al. 2013b).

## OVERVIEW OF PUBLISHED EXPLORATORY STRUCTURAL EQUATION MODELING APPLICATIONS WITH RELEVANCE TO CLINICAL AND SOCIAL SCIENCE RESEARCH

### Content Analysis of Published ESEM Applications

In this section we provide an overview of ESEM applications in clinical and psychological research (Table 2). We begin with a summary of a Google Scholar search on all ESEM references, starting with the first two publications, which appeared together in a dedicated issue of *Structural Equation Modeling*: the statistical background to ESEM with some simulated examples (Asparouhov & Muthén 2009), and the first published empirical application of ESEM (Marsh et al. 2009). We identified 103 full papers in the public domain, although only 91 were published journal articles. Because ESEM is a relatively new statistical strategy, the total number of citations of these 103 papers was 680, and this was dominated by citations to the first two publications (Asparouhov & Muthén 2009, 185 citations; Marsh et al. 2009, 101 citations). Not surprisingly, the number of publications has grown steadily over time, from 8 in 2009 to 12 in 2010, followed by 38 in 2012 and 23 in the first part of 2013.

Sixteen of 103 studies did not actually use ESEM (typically it was noted as a direction for future research to address limitations of the study), and another 18 studies only did traditional EFAs. We note that technically, EFA is a special case of ESEM, so that it is appropriate to label EFAs as ESEMs, but for the present purposes we distinguish between them. Another 13 studies extended the traditional ESEM approach by positing complex measurement error structures (e.g., correlated uniquenesses to test a priori method effects) that could not easily be incorporated into traditional EFAs. Nevertheless, this was usually done in combination with additional, more advanced ESEM applications (e.g., tests of a priori correlated uniqueness in longitudinal models). However, at least three studies used preliminary ESEMs for purposes of testing an initial factor structure, then reverted to the use of manifest scores for subsequent analyses (a strategy that is usually inappropriate).

Across the 103 articles (Table 2), particularly popular applications included tests of invariance across groups (34 studies) or occasions (15 studies). All 34 multigroup invariance studies began with tests of factorial invariance, but many went beyond this to include invariance of intercepts (strong measurement invariance, 31 studies) and further invariance constraints (strict measurement

**Table 2** Content analysis of ESEM articles, based on 103 publications

Publication	Frequency
Published journal article	91
Did not apply ESEM	16
EFA only	18
Switch to scale scores	3
Complex error	13
MIMIC	28
MG factor analysis invariance	34
MG:FL	34
MG:FL+int	31
MG:FL+uniq	22
MG:FL+var/covar	16
MG:FL+int+uniq	22
MG:FL+int+latent means	17
MG: full measurement invariance	12
LFA invariance	15
LFA:FL	15
LFA:FL+int	13
LFA:FL+uniq	12
LFA:FL+int+uniq	12
LFA:FL+int+uniq+latent means	9
LFA:FL+var/covar	7
LFA: full taxonomy	6
Other special features	8

Abbreviations: EFA, exploratory factor analysis; ESEM, exploratory structural equation model; FL, factor loading; Int, intercept; LFA, longitudinal factor analysis; MG, multigroup; MIMIC, multiple-indicator multiple-cause; uniq, uniqueness; var, variance.

invariance, 22 studies) in order to pursue tests based on latent means. Indeed, several studies (12) tested all 13 models in the Marsh et al. (2009) taxonomy of measurement invariance. Fifteen studies conducted similar invariance constraints across multiple occasions. We discuss below a few studies that integrated multiple groups and occasions into a single ESEM model.

Another popular ESEM strategy involved variations on the application of the basic MIMIC model. In some cases, the MIMIC model was used as an alternative to multigroup invariance tests to evaluate differential item functioning (see previous discussion). More generally, however, the MIMIC model was used to incorporate additional background variables or other constructs (latent or manifest) that were correlated with or regressed on the latent ESEM factors. Other distinctive or unusual applications of ESEM that demonstrate its flexibility are discussed further in the next section.

## Illustrative Applications Demonstrating Initial or Novel Applications of ESEM

The number and sophistication of ESEM studies have grown dramatically in just a short period of time. Here we present a history of selected research, demonstrating new and evolving ESEM strategies that have broad relevance to applied clinical and psychological research. We begin with

a summary of some of the earliest ESEM studies that first introduced key strategies, and we then discuss new or unique features of subsequent research that builds on these earlier studies.

**The first substantive ESEM study: student evaluations of teaching.** In the first empirical application of ESEM, Marsh et al. (2009) evaluated substantively important questions based on students' evaluations of university teaching using the multidimensional 36-item Students' Evaluations of Educational Quality (SEEQ) instrument. Although the a priori nine-factor solution was well supported by numerous EFAs (e.g., Marsh & Hocevar 1991), these findings were contested because CFA models failed to replicate these results (e.g., Toland & De Ayala 2005). Consistent with previous EFA research, Marsh et al. (2009) demonstrated that a well-defined ESEM structure fitted the data well, whereas the ICM-CFA models did not. Of critical importance, SEEQ factor correlations were substantially inflated in the CFAs [median (Md)  $r = 0.72$ ] compared to the ESEMs (Md  $r = 0.34$ ) in a way that undermined the discriminant validity and usefulness of SEEQ factors as diagnostic feedback. These two critical features of ESEM, compared to CFA/SEM, are common themes in many subsequent ESEM studies: the substantially improved fit and the substantially smaller correlations.

Based on their newly developed 13-model taxonomy of ESEM measurement invariance (Table 1), Marsh et al. (2009) used ESEM to test whether the SEEQ factor structure was fully invariant over the 13-year period that they considered; this was an important contribution to ESEM research and features in many subsequent ESEM studies (see Table 2). When year of administration was treated as a continuous variable, MIMIC ESEM models were also used to evaluate differential item functioning, and MIMIC ESEM growth models showed almost no linear or quadratic effects over this 13-year period.

MIMIC models also showed that relations with background variables (workload/difficulty, class size, prior subject interest, expected grades) were small in size and varied systematically for different SEEQ factors (e.g., class size was negatively related to the individual rapport factor but positively related to the organization factor), supporting the multidimensional perspective and a construct validity interpretation of the relations. Substantively important questions based on ESEM could not be appropriately addressed with either traditional approach (EFA or CFA). Together with the companion Asparouhov & Muthén (2009) article, the results of Marsh and colleagues set the stage for subsequent ESEM research.

**Big five personality.** Tests of the Big Five personality factor structures have been an active area of ESEM research; one that is particularly relevant to clinical and psychological sciences more generally. In a series of substantive-methodological synergies, Marsh and colleagues applied new and evolving ESEM methodology to Big Five personality responses. Marsh et al. (2010) used ESEM to resolve critical issues in Big Five factor structure for responses to the 60-item NEO-FFI instrument. Although supported by an impressive body of EFA research (see McCrae & Costa 1997), CFAs have failed to replicate these findings and have resulted in substantially inflated correlations relative to EFA results and Big Five theory.

The CFA results have led some methodologists (e.g., Vassend & Skrondal 1997) to question the factor structure of the NEO instruments—the most widely used Big Five personality instruments—whereas some Big Five substantive researchers have questioned the appropriateness of CFA for Big Five research (e.g., McCrae et al. 1996). Thus, McCrae et al. (1996, p. 568) 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.” However, rejecting the appropriateness of CFA for Big Five research would apparently mean forgoing the many advances in statistical methodology associated with CFA in personality



research, which would be an unfortunate state of affairs for a research area where factor analysis is so critical. Marsh et al. (2010) proposed ESEM to resolve these long-standing dilemmas in Big Five research, demonstrating that ESEM fitted the data better and resulted in substantially more differentiated (less correlated) factors than did CFA ( $Md r's = 0.20$  versus  $0.06$ ). They then applied the newly developed 13-model ESEM taxonomy of measurement invariance in relation to gender to establish the invariance of factor loadings, factor variances-covariances, item uniquenesses, correlated uniquenesses, and item intercepts, demonstrating with latent means that women score higher on all NEO Big Five factors.

Demonstrating the flexibility of ESEM, Marsh et al. (2010) proposed a complex structure of measurement errors to account for the fact that items from the 60-item NEO-FFI represented one of six subfacets representing each Big Five factor, based on the much longer 240-item NEO-Personality Inventory (PI) instrument from which it was derived. However, items on the FFI were not chosen in relation to facets, so some facets were overrepresented and others were not represented at all. Hence, Marsh, et al. (2010) treated the facets as method factors represented by correlated uniquenesses among items from the same facet. Although the introduction of this a priori error structure substantially improved the fit of both CFAs and ESEMs, the fit of the CFAs was still not adequate and was much poorer than that of ESEMs. In agreement with McCrae et al. (1996), Marsh et al. (2010) argued for the inappropriateness of the ICM-CFA factor structure for personality research but demonstrated that important strengths of the CFA approach could still be harnessed by applied researchers through the application of ESEM.

Furnham et al. (2013) used ESEM to evaluate the factor structure for the Big Five responses (the 240-item NEO-PI-R) based on a large ( $N = 13,234$ ) sample in a high-stakes job-related context. The NEO-PI-R is structured such that each of the 5 personality constructs is represented by 6 facets, and each facet is represented by 8 items (i.e.,  $5 \text{ factors} \times 6 \text{ facets} \times 8 \text{ items} = 240 \text{ items}$ ). Furnham et al. used 30 facet scale scores as the starting point of their analysis rather than the 240 items. Consistent with the Marsh et al. (2009) study of the 60-item NEO-FFI, they reported that ESEM fit the data substantially better than did CFA. Multigroup ESEMs showed support for strict (factor loading, intercept, uniqueness) invariance over gender. We note, however, that facet scores represent a special case of parcel scores, discussed previously (Marsh et al. 2013a).

Marsh et al. (2013b) used ESEM to test theoretical predictions about how Big Five factors vary across the lifespan with gender, age, and their interaction, based on the 15-item Big Five Inventory in the British Household Panel Survey ( $N = 14,021$ ; ages 15–99 years). ESEM fitted the data substantially better and resulted in much more differentiated (less correlated) factors than did CFA. Methodologically, they extended ESEM (first introducing ESEM-within-CFA models and a hybrid of multigroup and MIMIC models—see previous discussion), evaluating full measurement invariance and latent mean differences over age, gender, and their interaction. Substantial nonlinear age effects based on longitudinal ESEM models led to the rejection of the plaster hypothesis (that personality becomes set like plaster by age 30; Costa & McCrae 1994) and the maturity principle (that people with increasing maturity become more dominant, agreeable, conscientious, and emotionally stable; Caspi et al. 2005). However, the ESEM longitudinal results did support the newly proposed “la dolce vita effect”: that in later years, individuals become happier (more agreeable and less neurotic), more self-content and self-centered (less extroverted and open), more laid back and satisfied with what they have (less conscientious, open, outgoing, and extroverted), and less preoccupied with productivity.

In this same study, Marsh et al. (2013b) extended MIMIC ESEM strategies to tests of multigroup invariance, including tests specifically designed to investigate the loss of information due to categorizing continuous variables in multigroup approaches to invariance. First they conducted separate tests of measurement invariance over gender and over three age categories (young, middle,

old). Then they formed six groups, representing all combinations of two gender groups (male, female) and three age groups (young, middle, old), and tested measurement invariance across these six groups. To evaluate this multigroup invariance model, they introduced EwC (see previous discussion), which allowed them to partition latent mean differences into tests of age (linear and nonlinear), gender, and interaction effects. Finally, they extended the MIMIC/multiple-group hybrid approach by adding MIMIC age effects (linear and quadratic) to the gender-age multiple group models. In this way, they estimated the combined effects of age—based on continuous age (MIMIC) and multiple age categorical groups—and their interaction with gender.

**Bullying/victimization.** Marsh et al. (2011b) used ESEM to evaluate the responses to the 36-item, 6-factor Adolescent Peer Relations Instrument (verbal, social, and physical facets of bully and victim factors), noting that previous research had failed to identify well-differentiated facets. Although ESEM fitted the data only marginally better than did CFA, correlations among the three bully factors and among the three victim factors ranged from 0.72 to 0.84 for CFA but only 0.32 to 0.53 for the ESEM. The very high CFA factor correlations detracted substantially from the usefulness of responses for individual diagnosis and research purposes. Indeed, this study shows that even when goodness of fit for CFA models is apparently reasonable, there can still be substantial differences in the size of correlations among the multiple factors (for a similar observation, see Marsh et al. 2011a).

Marsh et al. (2011b) demonstrated strong measurement invariance of factor loadings and intercepts over gender, year in school, and time, but identified a different pattern of correlations among the factors for boys and girls—particularly in relation to the physical component of bullying and victim factors. MIMIC ESEM models demonstrated support for convergent and discriminant validity in relation to a wide variety of other fully latent constructs relevant to bullying research (e.g., depression, 11 components of self-concept, locus of control, coping styles, anger management, attitudes toward bullies and victims; a total of 32 constructs based on 168 items plus single-indicator constructs of linear and quadratic components of age, gender, and age-by-gender interactions).

ESEM MIMIC models of age and gender differences across the six latent bully/victim factors demonstrated the flexibility of the ESEM approach. Boys had much higher scores for the physical (bully and victim) subdomains and somewhat higher scores for the verbal subdomains, but they did not differ from girls for the social subdomain. Linear and quadratic year-in-school effects showed that all six latent factors tended to be lowest in year 7, increased in year 8, remained reasonably stable in years 9 and 10, and then declined in year 11. However, the increases with year in school were stronger for the bully factors than for the victim factors, and were stronger for the verbal factors than for the social or physical factors.

This study was apparently the first to apply autoregressive path models of causal ordering with ESEM latent factors. Not only were bully and victim factors positively correlated, but there was also evidence of reciprocal effects, such that each was a cause and an effect of the other (i.e., over time, victims become bullies and bullies become victims). ESEM MIMIC models showed that bullies and victims had similar patterns of results with most of the covariates, suggesting that they were more alike to each other than to students who were neither bullies nor victims.

**Passion.** Marsh et al. (2013c) used ESEM to test theoretical predictions from the dualistic model of passion and the two-factor (harmonious and obsessive passions) passion scale. ESEM fitted the data substantially better than did CFA and resulted in better differentiated (less correlated) factors. Originally developed in French, the passion instrument was subsequently translated into English. Although with CFA there is a well-developed approach to testing measurement invariance over translations, this was the first study to extend this approach to ESEM, and it demonstrated

support for invariance across all models in the 13-model taxonomy (see **Table 1**). Another interesting feature of this study is that participants were asked to identify and then to complete the passion scale items in relation to their activity areas. This idiographic approach to the passion scale assumes implicitly that the same set of items is equally appropriate across different areas of passion. Tests of invariance over five passion activity groups (leisure, sport, social, work, and education) indicated that the same set of items was appropriate for assessing passion across a wide variety of activities—a previously untested, implicit assumption that greatly enhances practical utility. On the basis of ESEM MIMIC models, Marsh et al. (2013c) found support for the convergent and discriminant validity of the harmonious and obsessive passion scales on a set of validity correlates: life satisfaction, rumination, conflict, time investment, activity liking and valuation, and perceiving the activity as a passion.

**Exploratory ESEM.** We have emphasized the use of ESEM as a confirmatory tool when there exists a well-defined a priori factor structure. However, ESEM is also valuable as an exploratory tool (see discussion by Morin et al. 2013), as demonstrated by Mora et al.’s (2011) study of clinical adherence to medical treatments in a sample of asthmatic patients and Myers et al.’s (2011) study of self-efficacy in coaches of youth sport teams. In each study, the authors noted that the factor structure was not well established, used a combination of fit and interpretability based on alternative models, positing varying numbers of factors to select a best model, and then tested invariance over time based on four waves of data (Mora et al. 2011) or over the coach’s gender (Myers et al. 2011). Maïano et al. (2013) also used an exploratory approach to ESEM on responses to the Eating Attitudes Test to clarify the factor structure and eliminate weak items. In each of these studies, the authors argued that an exploratory approach to ESEM, guided by substantive knowledge of the instrument, provided important new insights into the underlying factor structure.

**ESEM higher-order and bifactor models.** The traditional CFA approach to higher-order factor analysis is not readily available with the current operationalization of ESEM (Asparouhov & Muthén 2009), in that only broad restrictions can be placed on the latent correlation matrix (e.g., completely uncorrelated factors, or fully invariant factor correlations over multiple groups or occasions). Marsh et al. (2009) proposed several strategies to overcome these limitations. One alternative was a two-stage approach in which the latent correlation matrix of first-order factors was the basis of second-order factor analysis. Their suggestion was operationalized by Meleddu et al. (2012; also see Pettersson et al. 2012) to define a higher-order happiness factor based on the multidimensional Oxford Happiness Questionnaire for the measurement of psychological well-being. They concluded that “results support the idea that well-being is multidimensional and that the different dimensions form a single superfactor” (Meleddu et al. 2012, p. 183).

In an alternative approach, Marsh et al. (2009) used a set of global rating items to define a global factor in addition to nine specific factors of teaching effectiveness, but the fit of this model was poorer than that of the ESEM model, in which the global rating item loaded separately on each specific factor. Although this is not discussed by Marsh and colleagues, at least the rationale of this approach is similar to the bifactor model “rediscovered” by Reise (2012), which is a viable alternative to traditional higher-order CFA models. Although Reise focused mainly on bifactor CFA models, he also noted that exploratory bifactor modeling is greatly underused in applied research, and he provides preliminary support for an exploratory bifactor model with target rotation that allows items to load on multiple group factors.

Although we know of no studies that focus specifically on ESEM bifactor models, Pettersson et al. (2012) suggested that this approach might be more useful than the two-stage ESEM approach proposed by Marsh et al. (2009). More specifically, Pettersson and colleagues posed a particularly

novel ESEM approach to evaluating the nature of item wording effects (positively and negatively worded items) and higher-order personality factors based on Big Five responses. They began with the two-step approach, based on the latent correlation matrix among factors in the first step being used to define higher-order factors in the second step. However, the higher-order structure, which primarily reflected the valence of items, was not particularly satisfactory. Although they did not actually use the label, they instead used the ESEM bifactor model with target rotation proposed by Reise (2012; also see Marsh et al. 2013a) to model one general evaluative factor and five content-specific ESEM Big Five factors. Consistent with other research identifying problems with negatively worded items, Pettersson et al. (2012) found that many of the negatively evaluated items contained almost no descriptive variance. After they controlled for the general evaluative factor, the Big Five content-specific factors contained both positively and negatively valued items loading high and low on the same factors. For example, extraversion had positive loadings on positive traits (spontaneous, sociable, and expressive) but also on negative-valued traits (wild, gushy, and overbearing); it had negative loadings on negative-valued traits (timid, withdrawn, and restricted) but also on positive-valued items (cautious, private, and discreet). Pettersson et al. (2012) discussed alternative interpretations of the global evaluative factor (e.g., a response bias or method factor, a general self-esteem factor, or even an evolutionary selection factor) and other applications of ESEM. Hence, this appears to be the first published application of an ESEM bifactor model—even though Pettersson and colleagues did not identify their approach as such. Also, as suggested by Morin et al. (2013), the EwC approach would allow researchers to test a higher-order factor structure based on a first-order ESEM measurement model; however, we found no published applications of this approach.

**Multitrait–multimethod analysis: convergent and discriminant validity.** Campbell & Fiske's (1959) multitrait-multimethod (MTMM) paradigm is perhaps the most widely used construct validation design to assess convergent and discriminant validity, and it is a standard approach for evaluating psychological instruments. In the MTMM approach, construct validity is assessed by measuring multiple traits with multiple methods. In psychological measurement studies, the multiple traits typically refer to the a priori multiple factors that an instrument is designed to measure (e.g., the Big Five factors in personality research). They used the term “multiple methods” very broadly to refer to multiple tests or instruments, multiple methods of assessment, multiple raters, or multiple occasions.

Although the original Campbell-Fiske guidelines are still widely used to evaluate MTMM data, important problems with the guidelines when they are based on manifest scores are well known (see reviews by Marsh 1988, 1995; Marsh & Grayson 1995). Ironically, even in highly sophisticated CFA approaches to MTMM data, a single (manifest) scale score is typically used to represent each trait-method combination, but it is stronger to incorporate the multiple indicators explicitly into the MTMM design (e.g., Marsh 1993, Marsh & Grayson 1995, Marsh & Hocevar 1988). When multiple indicators are used to represent each scale, CFAs at the item level result in an MTMM matrix of latent correlations, thereby eliminating many of the objections to the Campbell-Fiske guidelines. However, compared to ESEM solutions, the overly restrictive ICM-CFA model typically provides a poorer fit and results in inflated correlations among different factors that are particularly critical in MTMM studies, resulting in substantially poorer discriminant validity. Hence, ESEM is well suited to the construction of latent MTMM correlation matrices that can then be evaluated in relation to the Campbell-Fiske guidelines.

Campbell & O'Connell (1967) specifically operationalized the multiple methods in their MTMM paradigm as multiple occasions. Several MTMM ESEM studies (e.g., Marsh et al. 2011a,b, 2013c) using this MTMM design provide particularly strong approaches to evaluating

discriminant validity. In each of these studies, ESEM fitted the data better than did ICM-CFAs. Importantly, the inflated correlations among CFA factors substantially undermine support for discriminant validity relative to the results based on the ESEM factors.

In a particularly relevant application of the ESEM MTMM approach, Burns et al. (2013) evaluated ratings by mothers, fathers, and teachers for 26 symptoms of attention deficit-hyperactivity disorder (ADHD) and oppositional defiant disorder (ODD) behaviors for large samples of Thai adolescents and Spanish children. Because of the categorical nature of the data, they used robust weighted least squares estimation in combination with the complex design option to control for the hierarchical nature of the data (students nested within teachers so that each teacher made ratings of many children). Preliminary ESEMs for each country demonstrated support for an a priori three-factor model for each method (source: mothers, fathers, and teachers) considered separately. Correlated uniquenesses were included a priori for responses to the same item by different sources, although this was only done for responses by mothers and fathers. For both countries there was good support for the invariance of factor loadings and thresholds across the three sources, but latent means for symptoms were systematically higher for mothers and fathers than for teachers. For the Spanish sample of children, there was good support for convergent and discriminant validity, although agreement was much stronger between the two parents, and the moderately high correlations among factors for teacher ratings (0.52–0.62) detracted from discriminant validity. For the Thai adolescent sample, there was reasonable support for convergent and discriminant validity for ratings by the two parents. However, for teacher ratings, there was only weak support for convergent validity and little or no support for discriminant validity. The authors suggested that differences between the two samples might reflect age differences, cultural differences, or differences in the translation of the symptoms.

Burns et al. (2013) suggested that a potential weakness in the ESEM MTMM approach was the inability to apply more advanced CFA models that provide indexes of latent trait and latent method effects, but we are not entirely in agreement with this suggestion. First, a more detailed application of the original Campbell-Fiske criteria would have been more diagnostically useful than traditional CFA MTMM models, particularly given that these models typically begin with manifest scale scores that would clearly be suboptimal, as demonstrated by Burns et al. (2013). Second, it is possible to apply more advanced models using a two-stage approach (based on the latent MTMM matrix estimated in the first stage) or the EwC approach, described previously (see related discussion of higher-order factors). An important direction for future research is to explore how effective these and other evolving ESEM strategies are in providing more complex models of MTMM data, and indeed if there are any real advantages to these more complex models relative to a detailed application of the original Campbell-Fiske criteria to latent ESEM MTMM correlation matrices.


**ESEMs of randomized controlled trials.** In applied research there remains a tendency for “correlational” studies to embrace latent variable models, although experimental interventions and randomized controlled trials (RCTs) continue to rely on manifest analyses. However, most limitations in the use of manifest variables in correlational studies also apply to RCT-type studies. Indeed, in RCT research there is sometimes a serious neglect of rigorous psychometric evaluation of outcomes measures—factor structure, construct validity, and invariance over time and groups. Here we briefly summarize two RCT ESEM studies.

Kushner et al. (2013) used ESEMs to evaluate the RCT results of a cognitive-behavioral therapy (CBT) intervention on internalizing psychopathology for alcohol-dependent patients relative to a control group trained in muscle relaxation. On the basis of results from six measures of internalizing symptoms, ESEM identified a two-factor solution at baseline and at four-month



follow-up consisting of distress (depression, trait anxiety, worry) and fear (panic, social anxiety, agoraphobia). A potential concern is the use of manifest test or scale scores as indicators without testing the structure of responses at the item level. Nevertheless, a series of ESEM invariance tests showed strong measurement invariance across all combinations of the two times (longitudinal invariance) and over experimental and comparison groups (multigroup invariance). Latent means from this fully invariant ESEM model showed that both groups improved over time but that the CBT group improved significantly more in terms of distress (but not fear) reduction. The authors emphasized that their ESEM approach provides a practical solution to modeling comorbidity in a clinical trial and is consistent with converging evidence pointing to the dimensional structure of internalizing psychopathology.

Lang et al. (2011) applied ESEM to the 15-item Big Five Inventory from the German Socioeconomic Panel Study ( $N = 19,351$ ; ages 18–90). However, unlike the Marsh et al. (2013b) study of age differences in personality structure, the focus of Lang and colleagues was on three randomly assigned data collection methods (assisted face-to-face interviewing, computer-assisted telephone interviewing, and a self-administered questionnaire). For young and middle-aged adults, ESEM models of the five-factor structure supported strict invariance (factor loadings, intercept, and uniquenesses) across the three administration methods, although openness latent means were higher for telephone interviews. For older adults, the factor structure was less robust for the telephone interview approach, possibly due to the higher cognitive demands of this approach. Over the five-year interval between the two data collections, self-administered surveys showed stronger test-retest correlations. The authors showed (follow the **Supplemental Material** link in the online version of this article or at <http://www.annualreviews.org/>) that factor variances were invariant across method groups for all three age groups and provided further information on measurement invariance on age-by-method group comparisons. Methodologically, this represents a particularly sophisticated ESEM RCT study—incorporating full measurement invariance and latent means over randomly assigned intervention groups (the three administration methods), age groups, and time (the five-year test-retest interval) for a very large nationally representative sample of adults—that can readily be extended to RCT clinical studies.

 **Supplemental Material**

## **DIRECTIONS FOR FUTURE DEVELOPMENT**

### **Using ESEM Factors in Subsequent Analyses: Manifest Scores and Factor Scores Versus Latent Correlation Matrices and/or Plausible Values**

Applied researchers sometimes use preliminary factor analyses (EFA, ESEM, or CFA/SEM) to test their a priori factor structure as a means of testing the construct validity of interpretations of the latent factors, but they then construct manifest scores (e.g., scale scores or factor scores) in subsequent analyses. Although the use of factor scores is preferable to the use of scale scores, because factor scores are more closely related to the underlying factor structure, neither approach is generally appropriate. In particular, both scale and factor scores are manifest scores that do not provide appropriate correction for measurement error, which is likely to substantially distort subsequent analyses based upon them. In a related discussion of problems associated with parcels, Marsh et al. (2013b) suggested that the use of plausible values might be a viable alternative. This approach is routinely used in large-scale educational databases such as the Program for International Student Assessment (Organ. Econ. Coop. Dev. 2007) and the Trends in International Mathematics and Science Study (Olson et al. 2008). To the extent that the set of plausible values represents uncertainty associated with measurement error and that this uncertainty is incorporated

into the factor model, the plausible values are likely to be a more attractive alternative to the use of manifest scale or factor scores.

### **Juxtaposing EFA, CFA, ESEM, and Bayesian Structural Equation Models?**

In discussions of the typically inappropriate use of parceling strategies, we have noted the dilemma of researchers who collect large number of items with modest sample sizes of participants. We have argued that the use of item parcels is usually inappropriate and have suggested ESEM as a more viable alternative. However, new and evolving Bayesian statistical procedures are also especially useful for the evaluation of complex factor structures with small  $N$ s, where maximum likelihood might not be appropriate. Thus, for example, the Bayesian SEM (BSEM) procedure in Mplus fits a factor model in which cross-loadings and correlated uniquenesses can take on nonzero values with informative priors based on the researcher's judgment. As emphasized by Muthén & Asparouhov (2012), this BSEM rationale is similar in many ways to the target rotation with ESEM demonstrated here, but it apparently overcomes potential limitations of ESEM, particularly when the model is large relative to the sample size. Hence the primary justification for the use of parcels is likely to be superseded with further development of BSEM. We view the two approaches as complementary: Increasing knowledge based on ESEM provides a basis for specifying priors in BSEM, but additional research that juxtaposes these approaches is needed (for further discussion, see Muthén & Asparouhov 2012). Nevertheless, particularly when  $N$  is small, BSEM estimates are heavily dependent on the analyst's beliefs, such that informative priors do not allow the estimates to differ substantially from expected values, so BSEM is not a panacea under these circumstances.

### **Recommendations for Applied Clinical Researchers**

We end our review of ESEM theory and research with a series of recommendations relating to how clinical researchers might go about specifying and testing such models. We emphasize that like rules of thumb, the appropriateness of these recommendations is context dependent. Most clearly, in preliminary analyses at the level of individual items, researchers should compare ESEM and ICM-CFA measurement models based on all the constructs to be considered. In these preliminary measurement models, researchers should simply allow constructs to be correlated, even if subsequent SEMs are tested, because these SEMs are either equivalent to or nested under the measurement models. If the fit and parameter estimates (e.g., latent factor correlations) for the ICM-CFA do not differ substantially from the corresponding ESEM, on the basis of parsimony researchers should retain the CFA model as the starting point for subsequent analyses. However, a growing body of research suggests that this will rarely be the case. If the ESEM fit (and interpretability) are acceptable and better than the CFA fit, researchers should retain the ESEM measurement model as the basis of subsequent analyses. If neither ICM-CFA nor ESEM models fit the data, or the ESEM model fits much better but does not result in an interpretable solution, researchers should explore alternative (ex post facto) solutions at the item level with appropriate caution, using the exploratory approach to ESEM.

A potentially serious limitation of ESEM is its lack of parsimony relative to CFA. Nevertheless, expedient compromises between parsimony and accuracy in applied research (e.g., the use of parcel, factor, or scale scores, or very short scales) when sample sizes are modest in relation to the number of items are likely to be biased under typical conditions and should be avoided unless the very restrictive assumptions upon which they are based are met. If even the full measurement model is too complex to fit at the item level, researchers might evaluate the factor structure of logically defined subsets of factors in relation to different subsets of factors (e.g., the multiple factors based on the instrument in relation to multiple factors based on each of the other instruments in a pairwise

strategy). Also, evolving Bayesian estimation procedures might make large models more tractable. Nevertheless, the onus is still on researchers to justify the appropriateness of their a priori (or ex post facto) measurement model at the item level before proceeding to more complex models.

### SUMMARY POINTS

1. CFA/SEMs have largely superseded EFAs, but CFA/SEMs are usually too restrictive to provide acceptable goodness of fit for most psychological instruments. ESEM, an overarching integration of the best aspects of CFA/SEMs and traditional EFAs, provides a viable option.
2. Due to misfit associated with overly restrictive measurement models with no cross-loadings, CFAs typically produce inflated factor correlations compared to ESEMs and to known population values for simulated data. This detracts from discriminant validity, undermines diagnostic usefulness, and results in complicated biases in more complex models.
3. For simulated data with cross-loadings, ESEM estimates of factor correlations are more accurate than CFA estimates and are generally accurate—but less parsimonious—even when there are no cross-loadings in the population-generating model.
4. ESEM incorporates traditional EFA and CFA/SEMs as special cases, so that nearly all models able to be fitted with CFA/SEM can be fitted with ESEM, without the limitations of the overly restrictive CFA/SEM measurement structure.
5. ESEMs are sufficiently flexible to include in a single model, various combinations of CFA factors, multiple sets of ESEM factors, manifest (MIMIC) variables, multigroup and longitudinal data, bifactor models, complex error structures, and a priori equality constraints to test, for example, full measurement invariance and differential item functioning.
6. The 13-model taxonomy of ESEM invariance tests incorporates traditional CFA/SEM (covariance structure) and item-response-theory (mean structure) approaches for factor/measurement invariance, which illustrates ESEM's remarkable flexibility.
7. ESEM is primarily a confirmatory tool, but like traditional EFAs (and even CFA/SEMs) it can be used with appropriate caution as an exploratory tool in a way that has many potential advantages over EFA, CFA/SEM, and even the presently evolving Bayesian approaches.
8. Applied researchers are recommended routinely to conduct preliminary analyses at the level of individual items, comparing ESEM and CFA measurement models based on all constructs to be considered in order to compare the suitability of CFA/SEMs and ESEMs for subsequent analyses.

### FUTURE ISSUES

1. ESEM-within-CFA (EwC) approaches have been proposed because some specific models that can be fitted in CFA/SEM are not currently available in ESEM (e.g., partial factor loading invariance, higher-order factors, some specific invariance constraints), but limitations in the EwC have not been fully explored.

2. Multilevel and mixture models cannot easily be fitted with the current Mplus version of ESEM. Although these limitations may be addressed in the future, alternative approaches at present include treating the within- and between-covariance matrices as separate sets of factors, or the EwC approach.
3. Like EFA, ESEM suffers from rotational indeterminacy in that different rotation strategies result in different solutions that all fit the data equally well. Target rotation seems to provide a stronger basis for testing a priori structures, but more research is needed to establish best practice.
4. ESEM multitrait multimethod (MTMM) analyses, compared to conventional CFA MTMM approaches, should substantially improve support for discriminant validity for many psychological instruments and should be incorporated into more general models; however, more research is needed to explore this potential.
5. ESEM's lack of parsimony is a potential limitation, particularly for large numbers of indicators and/or small sample sizes. Some compromise solutions (e.g., item parceling, very short scales, scale/factor scores) are questionable in most situations, but other possibilities (plausible values, Bayesian analyses) need further research.
6. Further research is needed, juxtaposing EFA, CFA/SEM, ESEM, and Bayesian SEM (BSEM). With target rotation, evolving BSEM approaches have a similar rationale to ESEM and thus they are complementary; however, more mathematical and empirical research is needed.

## DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

## ACKNOWLEDGMENTS

The authors thank Bengt Muthén and Tihomir Asparouhov as well as other coauthors of the ESEM studies cited in this review (including Gregory Arief Liem, Oliver Lüdtke, Christophe Maïano, Andrew Martin, Benjamin Nagengast, Alexander Robitzsch, and Ulrich Trautwein) for helpful comments at earlier stages of this research. This research was supported in part by a grant to the first three authors from the Australian Research Council (DP130102713).

## LITERATURE CITED

- Asparouhov T, Muthén B. 2009. Exploratory structural equation modeling. *Struct. Equ. Model.* 16:397–438
- Bandalos DL. 2008. Is parceling really necessary? A comparison of results from item parceling and categorical variable methodology. *Struct. Equ. Model.* 15:211–40
- 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
- Brown TA. 2006. *Confirmatory Factor Analysis for Applied Research*. New York: Guilford
- Browne MW. 2001. An overview of analytic rotation in exploratory factor analysis. *Multivar. Behav. Res.* 36:111–50

---

Presents mathematical/statistical basis of ESEM.

---

- Burns GL, 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
- 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? *Multivar. Behav. Res.* 2:409–26
- 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.* 2:134–39
- Church AT, Burke PJ. 1994. Exploratory and confirmatory tests of the Big 5 and Tellegen's three- and four-dimensional models. *J. Personal. Soc. Psychol.* 66:93–114
- Cohen J. 1968. Multiple regression as a general data-analytic system. *Psychol. Bull.* 70:426–43
- 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: Am. Psychol. Assoc.
- 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. *Multivar. Behav. Res.* 44:147–81
- Dolan CV, Oort FJ, Stoel RD, Wicherts JM. 2009. Testing measurement invariance in the target rotated multigroup exploratory factor model. *Struct. Equ. Model.* 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 Test. Psychom. Methodol. Appl. Psychol.* 14:3–25
- Howarth E. 1972. A factor analysis of selected markers for objective personality factors. *Multivariate Behav. Res.* 7:451–76
- Jöreskog KG. 1969. A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika* 34:183–202
- 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
- Jöreskog KG, Sörbom D. 1979. *Advances in Factor Analysis and Structural Equation Models*. New York: Univ. Press Am.
- Jöreskog K, Sörbom D. 1993. *LISREL 8: Structural Equation Modeling with the SIMPLIS Command Language*. Chicago, IL: Sci. Softw. Intl.
- 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
- 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
- Little TD, Cunningham WA, Shahar G, Widaman KF. 2002. To parcel or not to parcel: exploring the question and weighing the merits. *Struct. Equ. Model.* 9: 151–73

---

Applies ESEM to derive a latent MTMM matrix; overcomes many objections to the Campbell-Fiske criteria.

---



---

RCT clinical intervention that demonstrates usefulness of ESEM in delineating symptoms of co-occurring anxiety and alcohol disorders.

---



---

Applies ESEM to evaluate results of an RCT study.

---



**Resolves long-standing debate about the appropriateness of EFA and CFA for the NEO-FFI Big Five personality inventory.**

**Demonstrates inappropriateness of item parcels with real/simulated data unless ICM-CFAs at item level fit, as well as ESEMs.**

**Empirically demonstrates ESEM; introduces 13-model taxonomy of ESEM invariance over multiple groups and occasions as well as application of ESEM MIMIC.**

**Introduces EwC; extends the hybrid integration of MIMIC and multigroup approaches to invariance; posits the la dolce vita effect of personality change in old age.**

- 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 AJ, Lafranchi MC, Therme P. 2013. The Eating Attitudes Test-26 revisited using exploratory structural equation modeling. *J. Abnorm. Child Psychol.* 41:775–88
- Marsh HW. 1988. Multitrait multimethod analysis. In *Educational Research Methodology, Measurement and Evaluation: An International Handbook*, ed. JP Keeves, pp. 570–80. Oxford, UK: Pergamon
- 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. Model.* 1:5–34
- Marsh HW. 1995. The analysis of multitrait multimethod data. In *International Encyclopedia of Education*, ed. TH Husen, TN Postlethwaite, pp. 5125–28. Oxford, UK: Pergamon. 2nd ed.
- Marsh HW. 2007. Application of confirmatory factor analysis and structural equation modeling in sport/exercise psychology. In *Handbook of Sport Psychology*, ed. G Tenenbaum, RC Eklund, pp. 774–98. New York: Wiley. 3rd ed.
- Marsh HW, Grayson D. 1994. Longitudinal stability of latent means and individual differences: a unified approach. *Struct. Equ. Model.* 1:317–59
- Marsh HW, Grayson D. 1995. Latent variable models of multitrait-multimethod data. In *Structural Equation Modeling: Concepts, Issues, and Applications*, ed. RH Hoyle, pp. 177–98. Thousand Oaks, CA: Sage
- 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 McArdle, pp. 275–340. 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. *Multivar. Behav. Res.* 33:181–220
- Marsh HW, Hocevar D. 1988. A new, more powerful approach to multitrait-multimethod analyses: application of second-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. *Teach. Teach. Educ.* 7:9–18
- Marsh HW, Liem GAD, Martin AJ, Morin AJS, Nagengast B. 2011a. 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. 2013a. Why item parcels are (almost) never appropriate: Two wrongs do not make a right—camouflaging misspecification with item parcels in CFA models. *Psychol. Methods* 18:257–84
- 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. Model.* 16:439–76
- Marsh HW, Nagengast B, Morin AJS. 2013b. 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
- Marsh HW, Nagengast B, Morin AJS, Parada RH, Craven RG, Hamilton LR. 2011b. 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'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. 2013c. Passion: Does one scale fit all? Construct validity of two-factor passion scale and psychometric invariance over different activities and languages. *Psychol. Assess.* 25:796–809
- 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. Personal. 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. Personal. 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, et al. 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 modeling: an introduction. In *Structural Equation Modeling: A Second Course*, ed. GR Hancock, RO Mueller, pp. 395–436. Greenwich, CT: IAP. 2nd ed.**
- Muthén B, Asparouhov T. 2012. Bayesian structural equation modeling: a more flexible representation of substantive theory. *Psychol. Methods* 17:313–35
- Muthén BO. 2002. Beyond SEM: general latent variable modeling. *Behaviormetrika* 29:81–117
- 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. [http://www.statmodel.com/course\\_materials.shtml](http://www.statmodel.com/course_materials.shtml)
- 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
- Olson JF, Martin MO, Mullis IVS, eds. 2008. *TIMSS 2007 Technical Report*. Chestnut Hill, MA: TIMSS, PIRLS Intl. Study Cent.
- Organ. Econ. Coop. Dev. 2007. *PISA 2006: Science Competencies for Tomorrow's World*. Paris: Organ. Econ. Coop. Dev.
- Overall JE. 1974. Marker variable factor analysis: a regional principal axes solution. *Multivar. Behav. Res.* 9:149–64
- Pettersson E, Turkheimer E, Horn E, Menatti AR. 2012. The general factor of personality and evaluation. *Eur. J. Personal.* 26:292–302**
- Reise SP. 2012. The rediscovery of bifactor measurement models. *Multivar. Behav. Res.* 47:667–96
- Sass DA, Schmitt TA. 2010. A comparative investigation of rotation criteria within exploratory factor analysis. *Multivar. Behav. Res.* 45:1–33
- Sass DA, Smith PL. 2006. The effects of parceling unidimensional scales on structural parameter estimates in structural equation modeling. *Struct. Equ. Model.* 13:566–86
- 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
- Thurstone LL. 1947. *Multiple Factor Analysis*. Chicago: Univ. Chicago Press
- 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

---

Provides a comprehensive overview of ESEM and subsequent extensions (e.g., EwC) with examples (including Mplus syntax) based on simulated data.

---



---

Provides the first published application of the ESEM bifactor model, with target rotation as an alternative to higher-order factor models.

---

- 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. Personal.* 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. *Hum. Resour. Manag. Rev.* 18:233–42



# Contents

Advances in Cognitive Theory and Therapy: The Generic Cognitive Model <i>Aaron T. Beck and Emily A.P. Haigh</i> .....	1
The Cycle of Classification: DSM-I Through DSM-5 <i>Roger K. Blashfield, Jared W. Keeley, Elizabeth H. Flanagan, and Shannon R. Miles</i> .....	25
The Internship Imbalance in Professional Psychology: Current Status and Future Prospects <i>Robert L. Hatcher</i> .....	53
Exploratory Structural Equation Modeling: An Integration of the Best Features of Exploratory and Confirmatory Factor Analysis <i>Herbert W. Marsh, Alexandre J.S. Morin, Philip D. Parker, and Gurminder Kaur</i> .....	85
The Reliability of Clinical Diagnoses: State of the Art <i>Helena Chmura Kraemer</i> .....	111
Thin-Slice Judgments in the Clinical Context <i>Michael L. Slepian, Kathleen R. Bogart, and Nalini Ambady</i> .....	131
Attenuated Psychosis Syndrome: Ready for DSM-5.1? <i>P. Fusar-Poli, W.T. Carpenter, S.W. Woods, and T.H. McGlashan</i> .....	155
From Kanner to DSM-5: Autism as an Evolving Diagnostic Concept <i>Fred R. Volkmar and James C. McPartland</i> .....	193
Development of Clinical Practice Guidelines <i>Steven D. Hollon, Patricia A. Areán, Michelle G. Craske, Kermit A. Crawford, Daniel R. Kivlahan, Jeffrey J. Magnavita, Thomas H. Ollendick, Thomas L. Sexton, Bonnie Spring, Lynn F. Bufka, Daniel I. Galper, and Howard Kurtzman</i> .....	213
Overview of Meta-Analyses of the Prevention of Mental Health, Substance Use, and Conduct Problems <i>Irwin Sandler, Sharlene A. Wolchik, Gracelyn Cruden, Nicole E. Mabrer, Soyeon Ahn, Abnaelee Brincks, and C. Hendricks Brown</i> .....	243

Improving Care for Depression and Suicide Risk in Adolescents: Innovative Strategies for Bringing Treatments to Community Settings <i>Joan Rosenbaum Asarnow and Jeanne Miranda</i> .....	275
The Contribution of Cultural Competence to Evidence-Based Care for Ethnically Diverse Populations <i>Stanley J. Huey Jr., Jacqueline Lee Tilley, Eduardo O. Jones, and Caitlin A. Smith</i> .....	305
How to Use the New DSM-5 Somatic Symptom Disorder Diagnosis in Research and Practice: A Critical Evaluation and a Proposal for Modifications <i>Winfried Rief and Alexandra Martin</i> .....	339
Antidepressant Use in Pregnant and Postpartum Women <i>Kimberly A. Yonkers, Katherine A. Blackwell, Janis Glover, and Ariadna Forray</i> .....	369
Depression, Stress, and Anhedonia: Toward a Synthesis and Integrated Model <i>Diego A. Pizzagalli</i> .....	393
Excess Early Mortality in Schizophrenia <i>Thomas Munk Laursen, Merete Nordentoft, and Preben Bo Mortensen</i> .....	425
Antecedents of Personality Disorder in Childhood and Adolescence: Toward an Integrative Developmental Model <i>Filip De Fruyt and Barbara De Clercq</i> .....	449
The Role of the DSM-5 Personality Trait Model in Moving Toward a Quantitative and Empirically Based Approach to Classifying Personality and Psychopathology <i>Robert F. Krueger and Kristian E. Markon</i> .....	477
Early-Starting Conduct Problems: Intersection of Conduct Problems and Poverty <i>Daniel S. Shaw and Elizabeth C. Shelleby</i> .....	503
How to Understand Divergent Views on Bipolar Disorder in Youth <i>Gabrielle A. Carlson and Daniel N. Klein</i> .....	529
Impulsive and Compulsive Behaviors in Parkinson's Disease <i>B.B. Averbeck, S.S. O'Sullivan, and A. Djanshidian</i> .....	553
Emotional and Behavioral Symptoms in Neurodegenerative Disease: A Model for Studying the Neural Bases of Psychopathology <i>Robert W. Levenson, Virginia E. Sturm, and Claudia M. Haase</i> .....	581



Attention-Deficit/Hyperactivity Disorder and Risk of Substance Use Disorder: Developmental Considerations, Potential Pathways, and Opportunities for Research <i>Brooke S.G. Molina and William E. Pelham Jr.</i> .....	607
The Behavioral Economics of Substance Abuse Disorders: Reinforcement Pathologies and Their Repair <i>Warren K. Bickel, Matthew W. Johnson, Mikhail N. Koffarnus, James MacKillop, and James G. Murphy</i> .....	641
The Role of Sleep in Emotional Brain Function <i>Andrea N. Goldstein and Matthew P. Walker</i> .....	679
Justice Policy Reform for High-Risk Juveniles: Using Science to Achieve Large-Scale Crime Reduction <i>Jennifer L. Skeem, Elizabeth Scott, and Edward P. Mulvey</i> .....	709
Drug Approval and Drug Effectiveness <i>Glen I. Spielmans and Irving Kirsch</i> .....	741
Epidemiological, Neurobiological, and Genetic Clues to the Mechanisms Linking Cannabis Use to Risk for Nonaffective Psychosis <i>Ruud van Winkel and Rebecca Kuepper</i> .....	767

## Indexes

Cumulative Index of Contributing Authors, Volumes 1–10 .....	793
Cumulative Index of Articles Titles, Volumes 1–10 .....	797

## Errata

An online log of corrections to *Annual Review of Clinical Psychology* articles may be  
found at <http://www.annualreviews.org/errata/clinpsy>