

FIT INDEX SENSITIVITY IN MULTILEVEL STRUCTURAL EQUATION MODELING

BY

Aaron Boulton

Submitted to the graduate degree program in Psychology  
and the Graduate Faculty of the University of Kansas in partial fulfillment of the  
requirements for the degree of Master of Arts

---

Chairperson Kristopher J. Preacher, Ph.D.

---

Todd D. Little, Ph.D.

---

Wei Wu, Ph.D.

---

Paul E. Johnson, Ph.D.

Date Defended: July 20, 2011

The Thesis Committee for Aaron Boulton  
certifies that this is the approved version of the following thesis:

FIT INDEX SENSITIVITY IN MULTILEVEL STRUCTURAL EQUATION MODELING

---

Chairperson Kristopher J. Preacher, Ph.D.

Date approved: July 20, 2011

## Acknowledgements

I would like to thank the four members of my committee, Drs. Kristopher Preacher, Todd Little, Wei Wu, and Paul Johnson. Dr. Preacher initially fostered my interest in MSEM and has provided tremendous enthusiasm throughout all phases of this project. Without his support, this project would not have been possible. I am indebted to my co-advisor Dr. Little whose advice led to major structural changes in the methods employed. In my opinion, these changes both strengthened the study and demonstrated possible improvements to the Monte Carlo simulation paradigm that is currently used in quantitative psychology. Many thanks to Dr. Wu for her valuable feedback as well as her recent work on SEM model fit which inspired many aspects of this study. Finally, I am grateful to Dr. Johnson for his technical assistance on R programming as well as his eagerness to challenge commonly held assumptions which has led to an enriched understanding of my study.

## Abstract

Multilevel Structural Equation Modeling (MSEM) is used to estimate latent variable models in the presence of multilevel data. A key feature of MSEM is its ability to quantify the extent to which a hypothesized model fits the observed data. Several test statistics and so-called fit indices can be calculated in MSEM as is done in single-level structural equation modeling. Accordingly, problems associated with these measures in the single-level case may apply to the multilevel case and new complications may arise. Few studies, however, have examined the performance of fit indices in MSEM. Furthermore, recent findings suggest that evaluating fit at each level separately is advantageous to evaluating fit for the overall model. Therefore, the purpose of the present study was to evaluate the sensitivity of several fit indices to misspecification in the cluster-level model under varying multilevel data conditions including the intraclass correlation coefficient, sample size configuration, and severity of model misspecification. Furthermore, two methods of level-specific fit evaluation were compared. Results from a Monte Carlo simulation study suggest that fit indices are affected by the ICC of model indicators and sample size configurations in MSEM. With the exception of the SRMR, all fit indices were less sensitive to cluster-level model misspecification at low indicator ICCs, large overall sample sizes, and smaller numbers of clusters. Discrepancies in fit information between the two methods of level-specific fit were observed at low ICC values. Finally, two fit indices rarely used in SEM applications revealed desirable properties in certain simulation conditions. Implications of the simulation results are discussed and a program for implementing level-specific fit evaluation in the R statistical language is provided.

## TABLE OF CONTENTS

Introduction.....	8
PART I: Multilevel Structural Equation Modeling (MSEM).....	12
The Muthén & Asparouhov (2008) MSEM Framework.....	16
PART II: Model Fit Assessment in SEM and MSEM.....	21
Fit Index Sensitivity.....	24
Model Fit in MSEM.....	27
Level-Specific Fit.....	28
Segregating Approach.....	29
Partially-Saturated Approach.....	30
Research Questions.....	32
PART III: Methods.....	32
Data Generation and Analysis Models.....	33
Study Conditions.....	34
Dependent Variables & Analyses.....	39
PART IV: Results.....	40
Model Convergence.....	40
Fit Index Sensitivity.....	41
Loess Curves.....	41
Regression Analyses.....	46
PART V: Discussion.....	52
Limitations and Strengths.....	55
Future Directions.....	57

References.....	58
Footnotes.....	66
Tables.....	67
Figures.....	73
Appendix A: R Program for Segregating Approach Implementation.....	83

## List of Tables and Figures

Table 1. *The Chi-square Test Statistic and Other Approximate Fit Indices*

Table 2. *Population Parameter Values for Data Generation*

Table 3. *Convergence Rates for Simulation 1*

Table 4. *Convergence Rates for Simulation 2*

Table 5. *Regression Results for Simulation 1*

Table 6. *Regression Results for Simulation 2*

Figure 1. *Data Generation CFA Model*

Figure 2. *Data Analysis CFA Model*

Figure 3. *Loess Curves for  $\chi^2$*

Figure 4. *Loess Curves for RMSEA*

Figure 5. *Loess Curves for Within-group SRMR*

Figure 6. *Loess Curves for Between-group SRMR*

Figure 7. *Loess Curves for CFI*

Figure 8. *Loess Curves for TLI*

Figure 9. *Loess Curves for GFI\**

Figure 10. *Loess Curves for AGFI\**

## Sensitivity of Fit Indices in Multilevel Structural Equation Modeling

### Introduction

Nested data are pervasive in the social sciences. Nested data arise when the units of observation in a study are grouped in some way. For example, students are nested within classrooms, employees within departments, and patients within physicians. Another form of nesting occurs when repeated observations are made on the same unit over time. Nested data can arise naturally such as individuals nested within families or by design such as assigning individuals to record daily diary entries. Furthermore, a nested dataset can have an arbitrary number of levels. For example, repeated measures may be nested within students, who are nested within classrooms, which are further nested within schools, and so on.

There are some complications inherent in the analysis of nested data. Conceptually, one can make incorrect inferences if the analysis is restricted to a single level of the data. An *ecological fallacy* (Robinson, 1950) occurs when relationships among variables at the group-level are assumed to hold at the individual-level. Conversely, an *atomistic fallacy* (Diez-Roux, 1998) can be made by generalizing effects at the individual level to the group level. The direction of relationships also can change when collapsing groups from heterogeneous populations (*Simpson's paradox*; Simpson, 1951). Statistically, nested data imply that units within groups will respond more similarly than units between groups. This residual correlation violates the assumption of independent observations that underlie most parametric statistical procedures. Such a violation often will result in underestimated standard errors and thus a larger Type I error rate (identifying effects not actually present in the population) as well as biased parameter estimates (Hox, 1998).

Historically, researchers have analyzed nested data by either ignoring or controlling for group dependencies. If one ignores the nested structure then either aggregation or disaggregation is used. *Aggregation* refers to the analysis of units at the group level while ignoring information at the individual level by aggregating (i.e., averaging) scores within groups. Several problems are associated with this approach, such as committing an ecological fallacy, reduced statistical power, unreliable group-level information, and the incorrect weighting of groups during parameter estimation (Lüdtke et al., 2008; Preacher, Zyphur, & Zhang, 2010). *Disaggregation* involves analysis at the individual level while ignoring the nestedness and can result in biased test statistics, standard errors, and parameter estimates (Hox, 1998; Julian, 2001), confounding within and between-group relationships (Cronbach, 1976), and committing the atomistic fallacy.

To avoid these problems, others have attempted to control or correct for nestedness. Controlling for within-group dependence typically involves the inclusion of dummy-coded variables that represent group membership in the statistical model, often at the expense of parsimony. Corrections to standard errors have also been developed and implemented in several statistical software packages (Huber, 1967; White, 1982). Although such approaches can mitigate bias in parameter estimates and standard errors, information at other levels of the data is lost and generalizability is restricted only to the groups in a given sample.

A theme underlying each of these methods is that the dependence arising from nested data is superfluous to the intended analysis and must be either ignored or controlled for. More recently, however, investigators have begun to appreciate within-group dependence as a substantively interesting phenomenon that raises new and important questions (Hox, 1998). Multilevel modeling (MLM) is an extension of multiple regression analysis that provides a framework for such questions to be addressed (Raudenbush & Bryk, 2002; Snijders & Bosker,

1999). MLM, also known as hierarchical linear modeling, random coefficients modeling, or mixed modeling, allows for heterogeneity in regression parameters across groups. Different groups may have different mean levels (intercept) and conditional relationships (slopes) for a given outcome, and allowing these parameters to vary across groups addresses several of the aforementioned problems while providing new analytic insights. For example, MLM permits investigators to analyze cross-level interactions in order to determine whether within-group effects are conditional on group-level variables.

The key to MLM is the specification of intercept and slope parameters as random variables. Instead of estimating separate intercepts and slopes for each group as would be done using a fixed-effects approach, only a few parameters describing the distributions of the random effects are estimated. Fewer estimated parameters imply more parsimonious models, while at the same time generalizability is maximized given the specification of regression coefficients as random variables. In summary, MLM provides a concise and efficacious method for analyzing data at multiple levels of a hierarchy.

As is the case with any statistical procedure, MLM has its disadvantages. First, MLM assumes that variables have been measured without error, an assumption often violated in practice that can result in attenuated regression coefficients. Second, complex hypotheses involving multiple dependent variables are either difficult or impossible to test. Finally, MLM generally does not provide information regarding global model fit (see Wu, West, & Taylor, 2009, for an overview of model fit in MLM). As many have noted, however, the disadvantages of MLM are precisely the strengths of the structural equation modeling (SEM) framework (Bauer, 2003; Kline, 2011). SEM, also referred to as latent variable analysis, was developed out of the factor analytic tradition in psychology and involves modeling the relationships between

unobserved latent variables. SEM removes measurement error via explicit measurement models, complex theoretical models can be defined and tested, and several measures of global model fit are available. Likewise, the limitations of SEM are complemented by MLM. SEM assumes independent observations and therefore cannot account for or model group dependencies, a task that MLM was explicitly designed to do.

Clearly, these approaches have much to offer one another, and methodologists have worked to foster their synergy. The result is multilevel structural equation modeling (MSEM), a general analytic framework that combines the strengths of the SEM and MLM traditions (Kline, 2011; Muthén & Asparouhov, 2011). MSEM was conceptualized over four decades ago (Goldstein & McDonald, 1988; Hännqvist, 1978; Muthén, 1989; 1990; Schmidt, 1969), but only recently have analytical and computational advances made MSEM accessible to the larger research community. For those familiar with the MLM tradition, MSEM permits random intercepts and slopes such that cross-level interactions involving contextual variables can be estimated and evaluated. In contrast to MLM, however, outcomes need not be restricted to the lowest level of the data. For those familiar with the SEM tradition, MSEM is a latent variable technique, and thus measurement error in both predictor and outcome variables is controlled for. Additionally, MSEM provides measures of model fit and allows complex theoretical models to be specified and tested at multiple data levels.

MSEM is a fertile area for methodological inquiry. The extent to which findings from the MLM and SEM literatures hold in MSEM presents an interesting question that has largely been unexamined. In particular, the evaluation of model fit has been an active area of SEM research for the past 30 years. Little is known, however, about model fit assessment in MSEM. Popular measures of fit are able to be calculated for such models, yet it is unclear how the complexities

of multilevel data and new estimation methods influence their performance and sensitivity. The present study will employ a Monte Carlo simulation to address this general question. Before specific research questions are outlined, a review of the MSEM framework and model fit evaluation is warranted.

This paper proceeds as follows. Section I provides an overview of MSEM, for which many approaches now exist. Recent methods address the limitations of earlier developments (Ansari, Jedidi, & Jagpal, 2000; Muthén & Asparouhov, 2008; Rabe-Hesketh, Skrondal, & Pickles, 2004) and are currently implemented in accessible software. Of these, the general model of Muthén & Asparouhov (2008) is described in detail. Following this presentation, the discussion will turn to the issue of model fit in SEM (Section II). A summary of fit indices is provided, followed by a review of the extensive literature evaluating their performance. Section II ends with a discussion of fit in MSEM and current research questions. Section III outlines the Monte Carlo simulation used in this study. Recent insights into Monte Carlo investigations of fit index sensitivity are incorporated and highlighted. Results of the simulation are reported in Section IV and discussed in Section V.

#### PART I: Multilevel Structural Equation Modeling

The roots of MSEM can be traced to the dissertation of Schmidt (1969), in which a maximum likelihood (ML) estimator was developed for decomposing observed variables into latent sources of variation: between-cluster variation and within-cluster variation (Kaplan, Kim, & Kim, 2009). It follows then that a total population covariance matrix,  $\Sigma_T$ , can also be decomposed into the sum of a between-group covariance matrix,  $\Sigma_B$ , and a within-group covariance matrix,  $\Sigma_W$ . That is,

$$\Sigma_T = \Sigma_B + \Sigma_W \quad (1)$$

The sample estimate of  $\Sigma_W$ , referred to as the pooled within-group covariance matrix,  $S_{PW}$ , is calculated by

$$\mathbf{S}_{PW} = \frac{1}{N - J} \sum_{j=1}^J \sum_{i=1}^{n_j} (\mathbf{y}_{ij} - \bar{\mathbf{y}}_j)(\mathbf{y}_{ij} - \bar{\mathbf{y}}_j)' \quad (2)$$

where  $J$  is the total number of groups,  $n_j$  is the within-group sample size for cluster  $j$ ,  $N$  is the total sample size (i.e.  $\sum_{j=1}^J n_j$ ),  $\bar{\mathbf{y}}_j$  is a  $(p \times 1)$  vector of means for cluster  $j$  on  $p$  variables, and

$\bar{\mathbf{y}}_{ij}$  is a  $(p \times 1)$  vector of observed scores for individual  $i$  in cluster  $j$ . Similarly, the sample estimate of  $\Sigma_B$ , designated  $S_B$ , is calculated by

$$\mathbf{S}_B = \frac{1}{J - 1} \sum_{j=1}^J n_j (\bar{\mathbf{y}}_j - \bar{\mathbf{y}})(\bar{\mathbf{y}}_j - \bar{\mathbf{y}})' \quad (3)$$

where  $J$ ,  $n_j$ , and  $\bar{\mathbf{y}}_j$  are defined as before, and  $\bar{\mathbf{y}}$  is a  $(p \times 1)$  vector of overall sample means.

Muthén (1990) demonstrated that, for the case of balanced data in which  $n_j = n$  for all clusters  $j$ ,

$$\mathbf{S}^*_{PW} = \Sigma_W \quad (4)$$

and

$$\mathbf{S}^*_B = c\Sigma_B + \Sigma_W. \quad (5)$$

In Equation 5,  $c$  is a scaling parameter equal to  $\bar{n}_j$ . Muthén (1989, 1990) also demonstrated that separate models could be estimated for  $\mathbf{S}^*_{PW}$  and  $\mathbf{S}^*_B$  using the multiple-group option of standard SEM software and treating the two input matrices as “groups”. However, in the unbalanced case,  $\mathbf{S}^*_B$  is a biased estimator of the between-group covariance matrix as no single value of  $n$  applies to all clusters. In such a case, the following model holds for subsets of groups

where subset  $d$  consists of groups with an equal number of within-group units  $n_j$ ,

$$\mathbf{S}^*_{\mathbf{B}d} = c_d \boldsymbol{\Sigma}_{\mathbf{B}} + \boldsymbol{\Sigma}_{\mathbf{W}} . \quad (6)$$

Each subset  $d$  has a different scaling parameter,  $c_d$ . To estimate the model described in Equation 6 using standard SEM software, a separate between-group model for each subset  $d$  is estimated. Equality constraints are placed on all parameters and a mean structure is included across all between-group models (Muthén, 1990, 1994). It is not hard to imagine that programming and estimation for such a model can quickly become intractable.

To address this issue, Muthén (1989, 1990) proposed calculating a single  $\mathbf{S}^*_{\mathbf{B}}$  using an ad hoc estimator for the  $c$  scaling parameter, which is very close to the average sample size within clusters,

$$c^* = \frac{N^2 - \sum_j n_j^2}{N(J-1)} . \quad (7)$$

As a result of this formulation, Muthén (1989, 1990) developed a limited information maximum likelihood estimator, MUML, that approximates full-information ML estimates as the sample size at both levels becomes large (Hox, 1993; Hox & Maas, 2001; McDonald, 1994; Muthén, 1994; Yuan & Hayashi, 2005). This estimator has also been referred to as a *pseudobalanced* approach (McDonald, 1994). Goldstein (1987, 1995) proposed a different method for calculating  $\mathbf{S}^*_{\mathbf{P}\mathbf{W}}$  and  $\mathbf{S}^*_{\mathbf{B}}$  by “tricking” MLM software into estimating these quantities. Although this method addresses the issues of missing data and unbalanced clusters, the programming and data manipulation are cumbersome, and as  $\mathbf{S}^*_{\mathbf{P}\mathbf{W}}$  and  $\mathbf{S}^*_{\mathbf{B}}$  are not directly calculated but estimated, they are prone to sampling error (Hox & Maas, 2004).

The MUML approach has been used most frequently in the applied literature (Ryu, 2008). However, several MSEM methods have been developed based on a similar two-level covariance structure formulation (Lee, 1990; Lee & Poon, 1998; Liang & Bentler, 2004; McDonald, 1993; McDonald & Goldstein, 1989; Raudenbush, 1995). Although each approach differs with regards to computation and the ability to handle missing data or unbalanced clusters, each has been limited by their inability to estimate random slopes (Preacher et al. 2010). One solution for incorporating random slopes is to estimate certain SEM models within the MLM framework (Raudenbush, Rowan, & Kang, 1991). This also provides the advantage of accounting for an arbitrary number of data levels, but such models imposed overly restrictive measurement models and generally do not provide global fit information.

Recent computational and analytic advances have paved the way for more general MSEM formulations that address the aforementioned issues (Ansari et al., 2000; Jedidi & Ansari, 2001; Muthén & Asparouhov, 2008; Rabe-Hesketh et al., 2004). Anasari et al.'s (2000) framework, a Bayesian approach, has some theoretically interesting advantages as it does not rely on asymptotic theory, avoids high-dimensional integration, and can incorporate prior information into the estimation procedure. However, the method may be difficult for researchers to apply and their approach is not yet supported by generally available software. Conversely, the Muthén & Asparouhov (2008) approach is currently implemented in the SEM program *Mplus* (Muthén & Muthén, 1998-2010), and Rabe-Hesketh et al.'s (2004) method (called GLLAMM) is available via an add-on for the general statistical software package STATA.

Both methods can accommodate a variety of distributions for outcome variables (e.g., continuous, censored, binary, ordinal), adequately handle all forms of imbalance, including missing data, and accommodate random slopes. Rabe-Hesketh et al.'s (2004) GLLAMM

approach also allows for an arbitrary number of nested data levels, whereas the Muthén and Asparouhov's (2008) approach is often limited to two levels<sup>1</sup>. However, the GLLAMM approach is less computationally efficient for many models (Bauer, 2003; Preacher et al., 2010) and does not permit random slopes with latent covariates. Given the flexibility and computational efficiency of the Muthén and Asparouhov (2008) MSEM formulation, as well as its increasing use in the applied literature (e.g. Dedrick & Greenbaum, 2010; Purdy, Laschinger, Finegan, Kerr, & Olivera, 2010; Walsh, Matthews, Tuller, Parks, & McDonald, 2010), the following section will provide an overview of their method.

#### *Muthén & Asparouhov's (2008) MSEM framework*

The general MSEM framework as described by Muthén & Asparouhov (2008) is an extension of single-level SEM<sup>2</sup>. The measurement portion of the single-level SEM is

$$\mathbf{Y}_i = \mathbf{v} + \mathbf{\Lambda}\boldsymbol{\eta}_i + \mathbf{K}\mathbf{X}_i + \boldsymbol{\varepsilon}_i . \quad (8)$$

As before,  $i$  refers to an individual unit. In Equation 8,  $\mathbf{Y}_i$  is a  $(p \times 1)$  vector of observed scores on  $p$  variables,  $\mathbf{v}$  is a  $(p \times 1)$  vector of variable intercepts,  $\mathbf{\Lambda}$  is a  $(p \times m)$  matrix of factor loadings for  $m$  latent variables,  $\boldsymbol{\eta}_i$  is a  $(m \times 1)$  vector of the  $m$  latent variables,  $\mathbf{K}$  is a  $(p \times q)$  matrix of regression coefficients for the effects of  $q$  measured covariates on the  $p$  observed variables,  $\mathbf{X}_i$  is a  $(q \times 1)$  vector of observed scores on the  $q$  covariates, and  $\boldsymbol{\varepsilon}_i$  is a  $(p \times 1)$  vector of residual scores assumed to follow a multivariate normal distribution with zero means and a covariance matrix  $\boldsymbol{\Theta}$ . The measurement model shown here is also known as confirmatory factor analysis (CFA), and it is the part of an SEM model that expresses the observed variables as functions of underlying latent variables (the  $\boldsymbol{\eta}$ 's), observed covariates (the  $\mathbf{X}_i$ ), and residuals (the  $\boldsymbol{\varepsilon}_i$ 's).

The structural portion of the single-level SEM model specifies relationship patterns among the latent variables, which are allowed to covary in CFA models but not cause one

another. More formally, the latent variables are defined as functions of other latent variables, observed exogenous covariates, and residuals, as defined in Equation 9:

$$\boldsymbol{\eta}_i = \boldsymbol{\alpha} + \mathbf{B}\boldsymbol{\eta}_i + \boldsymbol{\Gamma}\mathbf{X}_i + \boldsymbol{\zeta}_i \quad (9)$$

where  $\boldsymbol{\eta}_i$  is defined as before,  $\boldsymbol{\alpha}$  is an  $(m \times 1)$  vector of intercepts,  $\mathbf{B}$  is an  $(m \times m)$  matrix of regression coefficients that specifies relationships among the latent variables,  $\boldsymbol{\Gamma}$  is an  $(m \times q)$  matrix of coefficients representing regressions onto covariates,  $\mathbf{X}_i$  is defined as before, and  $\boldsymbol{\zeta}_i$  is an  $(m \times 1)$  matrix of residual terms for the latent variables. The  $\boldsymbol{\zeta}_i$ 's are assumed to follow a multivariate normal distribution with means of zero and a covariance matrix  $\boldsymbol{\Psi}$ . Equations 8 and 9 imply

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{\Lambda}(\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Psi}((\mathbf{I} - \mathbf{B})^{-1})'\boldsymbol{\Lambda}' + \boldsymbol{\Theta} \quad (10)$$

$$\boldsymbol{\mu}(\boldsymbol{\theta}) = \mathbf{v} + \boldsymbol{\Lambda}(\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\alpha} \quad (11)$$

where  $\boldsymbol{\Sigma}(\boldsymbol{\theta})$  is the  $(p \times p)$  covariance matrix of the  $p$  observed variables in  $\mathbf{Y}_i$  expressed as a function of the parameters in the vector  $\boldsymbol{\theta}$ , and  $\boldsymbol{\mu}(\boldsymbol{\theta})$  is a  $(p \times 1)$  vector of means also expressed as a function of the parameters in  $\boldsymbol{\theta}$ . The quantities in 10 and 11 form the basis of the normal theory ML estimator, which is used to identify parameters in  $\boldsymbol{\theta}$  that minimize the discrepancy between the model-implied moments in  $\boldsymbol{\Sigma}(\boldsymbol{\theta})$  and  $\boldsymbol{\mu}(\boldsymbol{\theta})$  and the observed sample moments contained in the covariance matrix  $\mathbf{S}$  and the mean vector  $\boldsymbol{\mu}$ .

Equations 8 through 11 form the basis for the MSEM formulation given in Muthén and Asparouhov (2008). The key difference is that the parameter matrices are allowed to vary over clusters so as to define random effects at the between-group level. As such, Equations 12 through 14 represent the general MSEM model,

$$\mathbf{Y}_{ij} = \mathbf{v}_j + \boldsymbol{\Lambda}_j\boldsymbol{\eta}_{ij} + \mathbf{K}_j\mathbf{X}_{ij} + \boldsymbol{\varepsilon}_{ij} \quad (12)$$

$$\boldsymbol{\eta}_{ij} = \boldsymbol{\alpha}_j + \mathbf{B}_j \boldsymbol{\eta}_{ij} + \boldsymbol{\Gamma}_j \mathbf{X}_{ij} + \boldsymbol{\zeta}_{ij} \quad (13)$$

$$\boldsymbol{\eta}_j = \boldsymbol{\mu} + \boldsymbol{\beta} \boldsymbol{\eta}_j + \boldsymbol{\gamma} \mathbf{X}_j + \boldsymbol{\zeta}_j \quad (14)$$

where Equation 12 is a measurement model, Equation 13 is the within-group structural model, and Equation 14 is the between-group structural model. As is the case in the single-level formulation, the residual terms  $\boldsymbol{\varepsilon}_{ij}$  and  $\boldsymbol{\zeta}_{ij}$  are assumed to be multivariate normally distributed with means of zero and covariance matrices  $\boldsymbol{\Theta}$  and  $\boldsymbol{\Psi}$ .

Notice that Equations 12 and 13 are identical to the single-level SEM equations (8 and 9) with the addition of a  $j$  subscript on each term. Consequently, the terms in these equations are defined as before, with the exception that they are now allowed to vary at the between-group level. More precisely, the matrices containing the model parameters ( $\mathbf{v}_j$ ,  $\boldsymbol{\alpha}_j$ ,  $\boldsymbol{\Lambda}_j$ ,  $\mathbf{B}_j$ ,  $\mathbf{K}_j$ , and  $\boldsymbol{\Gamma}_j$ ) can vary at the between-group level, and the matrices containing the variable scores (both observed,  $\mathbf{Y}_{ij}$  and  $\mathbf{X}_{ij}$ , and unobserved,  $\mathbf{X}_{ij}$ ,  $\boldsymbol{\varepsilon}_{ij}$ , and  $\boldsymbol{\zeta}_{ij}$ ) can vary at both levels. Additionally, the elements within the latter matrices can be strictly within-group variables, strictly between-group variables, or variables with variance at both levels.

Equation 14 is less obvious to decipher. The vector  $\boldsymbol{\eta}_i$  is actually a stacked vector containing all  $r$  random effects from the parameter matrices  $\mathbf{v}_j$ ,  $\boldsymbol{\alpha}_j$ ,  $\boldsymbol{\Lambda}_j$ ,  $\mathbf{K}_j$ ,  $\mathbf{B}_j$ , and  $\boldsymbol{\Gamma}_j$ . That is,

$$\boldsymbol{\eta}_j = \begin{bmatrix} \text{vec}\{\mathbf{v}_j\} \\ \text{vec}\{\boldsymbol{\alpha}_j\} \\ \text{vec}\{\boldsymbol{\Lambda}_j\} \\ \text{vec}\{\mathbf{K}_j\} \\ \text{vec}\{\mathbf{B}_j\} \\ \text{vec}\{\boldsymbol{\Gamma}_j\} \end{bmatrix} \quad (15)$$

where  $\text{vec}\{\}$  is an operator that places the elements of its argument into a column vector. The vector  $\boldsymbol{\eta}_i$  is then defined as a function of the following elements:  $\boldsymbol{\mu}$ , an  $(r \times 1)$  vector of fixed effects (the means of the random effect distributions and structural intercepts);  $\boldsymbol{\beta}$ , an  $(r \times r)$  matrix of structural regression coefficients defining the relationships among the random effects;  $\boldsymbol{\gamma}$ , an  $(r \times s)$  matrix of regression coefficients for the random effects in  $\boldsymbol{\eta}_i$  regressed onto  $s$  between-group exogenous covariates contained in the  $(s \times 1)$  vector  $\mathbf{X}_j$  (which itself is an element in the partitioned matrix  $\mathbf{X}_{ij}$ ); and finally,  $\zeta_j$ , an  $(r \times 1)$  vector of residual scores with means of zeros and covariance matrix  $\boldsymbol{\psi}$ .

A notable characteristic of the model described in Equations 12, 13, and 14 is that standard analyses commonly used in the social and behavioral sciences are actually special cases of this model. That is, multiple regression analysis, path analysis, CFA, SEM, and two-level MLM are all estimable under this model. For example, consider a single-level CFA model in which neither the observed variables nor the latent variables are functions of exogenous predictors. Without any observed covariates, all equation terms containing the  $\mathbf{X}_{ij}$  matrix are removed. Furthermore, as CFA models do not specify causal paths among latent variables, Equation 13 is ignored. Finally, as the model is only estimated at the within-group level, all  $j$  subscripts are removed from the elements in Equation 12. Thus, we have the following model:

$$\mathbf{Y}_i = \mathbf{v} + \boldsymbol{\Lambda}\boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i \quad (16)$$

which implies

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{\Lambda}\boldsymbol{\Psi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta} \quad (17)$$

$$\boldsymbol{\mu}(\boldsymbol{\theta}) = \mathbf{v} + \boldsymbol{\Lambda}\boldsymbol{\alpha} \quad (18)$$

Equations 17 and 18 represent the standard CFA model-implied covariance matrix and mean vector (Brown, 2006).

In *Mplus*, the general MSEM model described in Equations 12 through 14 is estimated using a full-information maximum likelihood estimator (FIML, cf. Mehta & Neale, 2005). By default, a robust  $\chi^2$  statistic is calculated (Yuan & Bentler, 1998) and standard errors are estimated using Huber-White sandwich estimators (Hox, Maas, & Brinkhuis, 2010). In tandem, these will provide unbiased estimates under moderate violations of distributional assumptions. Additionally, a diagonally-weighted least squares (DWLS) estimator is available for model estimation with categorical dependent variables (Asparouhov & Muthén, 2007). In such instances, ML solutions require high-dimensional numerical integration which can often result in convergence problems, imprecise estimates, and severe computational burden. Estimation with the DWLS estimator reduces the number of dimensions by dividing a full model into multiple simple models that require only one- or two-dimensional integration. The DWLS implies robust  $\chi^2$  values and standard errors (Hox et al., 2010).

To summarize, early MSEM methods often require tenuous model and data assumptions (e.g., balanced clusters, no missing data), can be difficult to program, and cannot accommodate random slopes. Random slopes (and therefore cross-level interactions) are of key interest to users of MLM, and thus for a full synergy to exist between the SEM and MLM frameworks, an advantage of one approach should not be left out. With the development of FIML estimation for MSEM (Mehta & Neale, 2005) as well as advances in specialized (*Mplus*) and general statistical software (STATA), the full potential of SEM and MLM integration is now available. For applied researchers, new tools inspire the refinement of existing theories and the creation of new ones, or, to put it simply, scientific progress. For methodologists, however, new tools imply questions concerning its use, utility, and generalizability. As such, the discussion will now turn to a relatively unexplored topic in relation to MSEM: model fit evaluation. The evaluation of model

fit is a strength of the SEM framework that permits researchers to somewhat quantify the adequacy of a theoretical model as applied to real data. The following section will briefly review model fit as conceptualized in SEM. This review will provide a foundation from which the literature concerning the performance fit measures in SEM, and to a lesser extent, MSEM, can be understood.

## PART II: Model fit assessment in SEM and MSEM

### *Model Fit Statistics and Indices*

The goal of SEM is to recreate the means, variances, and covariances of multivariate sample data with a theoretical model. Formally, given a sample covariance matrix  $\mathbf{S}$  and mean vector  $\bar{\mathbf{y}}$ , one must specify and estimate a model (i.e., a system of linear equations) with parameter vector  $\boldsymbol{\theta}$  such that the covariance matrix,  $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ , and mean vector,  $\boldsymbol{\mu}(\boldsymbol{\theta})$ , implied by the model resemble  $\mathbf{S}$  and  $\bar{\mathbf{y}}$  as close as possible. Conceptually, SEM is similar to a simple regression analysis in which the goal is to minimize the distance between a set of observed scores (the sample data) and a set of predicted scores (the model-implied data). The only difference is that in the regression example we are trying to predict raw scores on a *single* dependent variable, whereas in SEM we are trying to predict elements in a matrix that represent the relationships between *several* variables, as well as the average level and spread of those variables.

It is well known in the SEM community that a statistical test of model fit is provided by transforming the discrepancy function value calculated during model estimation. Specifically, let  $F_{ML}$  represent a variable that is a function of the discrepancy between the sample data elements in  $\mathbf{S}$  and  $\bar{\mathbf{y}}$  and the model-implied elements in  $\boldsymbol{\Sigma}(\boldsymbol{\theta})$  and  $\boldsymbol{\mu}(\boldsymbol{\theta})$ . Further let  $\hat{F}_{ML}$  represent the

value obtained by minimizing such a function. Given multivariate normality, a large sample size, and proper model specification, the quantity

$$\mathbf{T}_{ML} = (N - 1)\hat{\mathbf{F}}_{ML} \quad (18)$$

is distributed as  $\chi^2_d$  with degrees of freedom  $d$  given by:  $d = \{[p(p + 1)/2] + p\} - t$ . As before,  $N$  is the total sample size and  $p$  is equal to the number of observed variables. The value  $t$  is the number of freely estimated parameters in  $\boldsymbol{\theta}$ .  $\mathbf{T}_{ML}$  provides a statistical test for the null hypothesis that the model perfectly reproduces the sample means and variances/covariances, or rather,  $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \mathbf{S}$  and  $\boldsymbol{\mu}(\boldsymbol{\theta}) = \bar{\mathbf{y}}$ . If the value obtained in Equation 18 is greater than a  $\chi^2$  variate corresponding to a pre-specified Type I error rate (usually .05), then the null hypothesis of perfect or exact fit is rejected. This test is actually a special case of the likelihood ratio (LR) test. The LR test obtains the difference in discrepancy values between two models—one of which is nested within the other (i.e. has a subset of that model's parameters)—and refers this value to a  $\chi^2$  distribution with degrees of freedom equal to the difference in degrees of freedom between models. In the SEM case, the reference model is a saturated model that simply estimates the elements in the sample covariance matrix and mean vector, and thus has a discrepancy value of zero and no degrees of freedom.

The  $\chi^2$  test, although intuitively appealing, has two interrelated flaws. As with other statistical tests, the power of the  $\chi^2$  test is influenced by sample size. Thus, at arbitrarily large sample sizes, every model will be rejected, even if the amount of misspecification is trivial and the model is of practical value (Bentler & Bonett, 1980). Conversely, at very small sample sizes, the test will not have enough power to detect even severe misspecifications. The second is perhaps more troubling as it encourages the use of small samples which yields imprecise parameter estimates and may violate assumptions underlying the test itself (Taylor, 2008).

Several measures of model fit, known as approximate fit indices (AFI) or practical fit indices, have been developed to avoid the problems associated with the  $\chi^2$  test. As West, Taylor, and Wu (in press) quip, “The decade of 1980s was the heyday of the development of new fit indices, and with apologies to songwriter Paul Simon--there must be 50 ways to index your model’s fit” (p. 7). Indeed, the number of AFIs is daunting, and methodologists have spent considerable effort to determine the performance and sensitivity of each under a variety of conditions. Before the results of these efforts are discussed, it is informative to classify AFIs according to certain characteristics. Formulas, sources, and other information for some AFIs are provided in Table 1.

Perhaps the most useful distinction between AFIs is that of absolute fit indices versus relative fit indices. *Absolute* fit indices quantify the overall fit of the model in terms of residuals, or the difference between  $\mathbf{S}/\bar{\mathbf{y}}$  and  $\mathbf{\Sigma}(\boldsymbol{\theta})/\boldsymbol{\mu}(\boldsymbol{\theta})$ . Many of these indices, such as the  $\chi^2/df$  ratio (Jöreskog, 1969), the root mean square error of approximation (RMSEA; Steiger, 1990; Steiger & Lind, 1980), and gamma hat (Maiti & Mukherjee, 1991; Steiger, 1989) are defined through  $T_{ML}$ . These fit indices can also be thought of as a comparison between the target model and the saturated model, which, as previously mentioned, perfectly reproduces the sample data. Other absolute fit indices such as the root mean square residual (RMR; Jöreskog & Sörbom, 1981) and standardized root mean square residual (SRMR; Bentler, 1995) are directly calculated from the residuals and do not depend on  $T_{ML}$ . *Relative* fit indices, such as the comparative fit index (CFI; Bentler, 1990), the Tucker-Lewis index (TLI; Tucker & Lewis, 1973), the relative noncentrality index (RNI; McDonald & Marsh, 1990), and the incremental fit index (IFI; Bollen, 1989), are calculated based on the comparison of a target model’s  $\chi^2$  to that of a worse-fitting model, known as the independence or null model. Conceptually, these indices can be thought of as indicating how much information that is lost by fitting the null model is recovered by estimating the target model.

Another distinction can be made between goodness of fit versus badness of fit indices (Taylor, 2008). Increases in goodness of fit indices reflect better model fit, whereas increases in badness of fit indices reflect worse fit. All comparative indices are goodness of fit indices, although not all absolute indices are badness of fit indices. Goodness of fit indices range between 0 and 1, with values closer to 1 indicating better fit. Some of these indices can exceed 1 in certain instances (West, Taylor, & Wu, in press). Badness of fit indices typically range from 0 to  $\infty$ , with 0 indicating perfect fit. It should be noted that the hypothesis of exact fit in the population is unrealistic—all models do not fit perfectly in the population—and thus other fit-based hypothesis tests using more realistic values have been proposed (MacCallum, Browne, & Sugawara, 1996).

Other classifications for fit indices have been proposed, although these will not be discussed (Sun, 2005; Tanaka, 1993). One final point worth considering, however, is that AFIs are *indexes* of model fit, and not statistical *tests* of model fit<sup>3</sup>. As will be discussed in more detail in the next section, several cutoff values have been proposed that are intended to reflect the difference between acceptable and unacceptable model fit. However, these are not based on the distributional properties of the indices—many of which are unknown—and thus cannot be used to make probabilistic inferences regarding a model. In the next section, the literature pertaining to fit index sensitivity will be reviewed. Given that AFIs were designed to detect model misspecification without undue influence of sample size or other extraneous factors, methodologists have been keenly interested in how indices perform under a variety of data and analytic conditions.

### *Fit Index Sensitivity*

Early efforts to evaluate the performance of fit indices were focused in both their substance and method. Given the problems previously mentioned with the  $\chi^2$  test, methodologists were most interested in whether any of the newly developed measures were sensitive to sample size or estimation method. To answer these questions, properly specified CFA or SEM models were fit to data at varying sample sizes and thus any observed variation in fit indices was attributed to the specific study manipulations (Anderson & Gerbing, 1984; Bearden, Subhash, & Teel, 1982; Ding, Velicer, & Harlow, 1995; Sugawara & MacCallum, 1993). Key findings from these early studies include: (1) Sample size still had non-trivial effects on some AFIs (notably, the NFI), and (2) Fit indices varied, some considerably, as a result of different estimation methods (maximum likelihood versus generalized least squares).

Eventually, researchers became interested not only in the performance of fit indices for correctly specified models, but also for misspecified models (Browne & Cudeck, 1993). Specifically, the question was whether variance in fit indices was attributable to model misspecification (a desirable outcome) or the specific data and modeling conditions under which estimation took place (an undesirable outcome; e.g. sample size, non-normality, number of variables). Although refined in various ways, this approach toward evaluating fit index sensitivity remains popular and has been implemented in a number of Monte Carlo simulations over the last 20 years (Fan & Sivo, 2005; 2007; Fan, Thompson, & Wang, 1999; Hu & Bentler, 1998; 1999; La Du & Tanaka, 1989; Marsh, Balla, & Hau, 1996; Marsh, Balla, & McDonald, 1988; Nevitt & Hancock, 2000; Olsson, Foss, Troye, & Howell, 2000; Olsson, Troye, & Howell, 1999).

Of particular note in the literature on fit index sensitivity is the work of Hu & Bentler (1998; 1999). Two main outcomes resulted from these studies. First, the authors found that the

SRMR was more sensitive to structural misspecification compared to other absolute and comparative indices, which were more sensitive than the SRMR to measurement misspecification. This led the authors to recommend a two-index strategy for model fit evaluation in which the SRMR as well as another well-performing index are considered. Second, the authors investigated a variety of cutoff values for several AFIs and, based on minimizing Type I and Type II error rates, proposed new and more stringent cutoff values by which models could be considered to demonstrate acceptable fit.

Several criticisms of these new values—as well as the use of cutoff values in general—have been raised by methodologists since their publication (Chen, Curran, Bollen, Kirby, & Paxton, 2008; Fan & Sivo, 2007; Marsh, Hau, & Wen, 2004; Yuan, 2005). The primary argument against the use of such values is that factors that should be unrelated to the fit of a hypothesized model still influence some indices, and thus appropriate cutoff values will differ according to the characteristics of the data and the model of interest. In simulation studies, factors such as sample size (Chen et al., 2008; Fan & Sivo, 2005; Taylor, 2008), the number of variables, (Breivik, & Olsson, 2001; Kenny & McCoach, 2003), the type of model (Fan & Sivo, 2007), and indicator reliabilities (Beauducel & Wittmann, 2005) have all accounted for variance in fit indices examined above and beyond actual misspecification in the model. Interestingly, even Hu & Bentler (1999) cautioned against dogmatic use of such values without a careful consideration of other aspects of the model as well as its substantive implications.

While the debate over the use of statistical tests of fit or cutoff values continues, new avenues of fit index sensitivity are currently being explored. Some authors have emphasized that the severity of model misspecification should be more explicitly measured and included as a design factor in Monte Carlo simulations (Fan & Sivo, 2005; Taylor, 2008; Wu & West, 2010).

If included, methodologists could determine whether fit indices perform differentially over different levels of misspecification, and the interpretation of other design factors will not be confounded by the severity of misspecification. For example, Fan and Sivo (2005) found that Hu and Bentler (1998) confounded type of misspecification with severity of misspecification, and replications in which severity was controlled suggested that the two-index strategy was unsupported. Also, index sensitivity to misspecification in mean structures is currently being explored (Leite & Stapleton, 2006; Wu, Taylor, & West, 2009; Wu & West, 2010).

The present study builds on previous work in this area by examining fit index sensitivity in MSEM. Given the structure of nested data, assessing model fit in MSEM poses additional complications that have not been fully explored or discussed in the methodological literature. In the following section, model fit evaluation in MSEM will be reviewed.

#### *Model Fit in MSEM*

A few simulation studies have investigated model fit in MSEM. Julian (2001) considered the consequence of estimating a single-level CFA with nested data. A four-factor structure accounting for the variation among 16 variables was generated at the within-group level, while three different factor structures (2, 4, 5 factors) were generated at the between-group level, and subsequently ignored in the analysis. Only the  $\chi^2$  test statistic was examined. Results showed that the  $\chi^2$  statistic was positively biased for higher levels of the ICC. Furthermore, greater inflation was observed as the group/member ratio decreased. Finally, there was no effect of the different between-group model structures on the  $\chi^2$  test statistic.

Hox and Maas (2001) examined the performance of the MUML estimator in estimating a multilevel CFA under differing levels of ICC, number of groups, group size, and group imbalance. Results suggested that small positive biases occurred with higher ICCs and greater

group imbalance for the  $\chi^2$  test statistic. However, in a later study the authors questioned these results, pointing out that the ICC condition was confounded with the amount of systematic variance specified in the between-group model (Hox et al., 2010). This latter study also investigated the performance of MUML, in addition to variations of the FIML and DWLS estimators, under similar conditions. It was found that the FIML estimator without robust  $\chi^2$ /standard errors and a DWLS estimator with mean- and variance- corrected  $\chi^2$ /standard errors produced the most accurate  $\chi^2$  values. The MUML-produced  $\chi^2$  values were most biased, and interestingly, the bias increased as the average within-group size increased. This is in contrast to the analytical findings of Yuan and Hayashi (2005) who demonstrated that small within-group sample sizes contribute to a higher coefficient of variation, which they suggest is the main determinant of the bias associated with MUML.

There are a few points to consider here. First, only correctly specified models under varying sample sizes, estimators, and ICCs were estimated in these studies. This focus mirrors that of the early simulation studies on fit index sensitivity in correctly-specified SEM models. Not surprisingly, it will be important to investigate the effects of these various design factors in misspecified models in the future. Second, only the normal-theory  $\chi^2$  test statistic was examined. Other test-statistics based on discrepancy functions exist for MSEM (Yuan & Bentler, 2007), and furthermore, AFIs were not investigated. Finally, *simultaneous* estimation was used for the analysis models. Simultaneous estimation refers to the fact that parameters for both the within-group and between-group models are estimated within a single discrepancy function. This approach can complicate the assessment of model fit in MSEM, as described in the next section.

#### *Level-Specific Fit in MSEM*

In routine applications of MSEM, estimation occurs simultaneously at each level, and thus the likelihood discrepancy value is a combination of misfit at each level of the data. There are several disadvantages to model fit evaluation based on this combined discrepancy function value (Ryu & West, 2009; Yuan & Bentler, 2007). First, it is unclear at which level misspecification is occurring. Second, the discrepancy value is unequally weighted by the larger unit-level sample size, and thus fit indices may not be sensitive to cluster-level misfit. Third, power is reduced for model fit test statistics. Finally, misspecification at one level can affect parameter estimates at the other levels. Motivated by these issues, two methods have been proposed as alternatives to the evaluation of model fit under simultaneous estimation and are described next.

*Segregating Approach.* Yuan and Bentler (2007) proposed estimating each level of the model separately and assessing fit for each model as in single-level SEM. This approach has been referred to elsewhere in the literature as the *segregating approach* (Ryu & West, 2009). In their study, the authors developed an estimator of the pooled within-group and between-group covariance matrices, as well as the asymptotic covariance matrices of these estimates. To use their approach, one first estimates the two covariance matrices and then uses the estimates as input data for single-level analyses in conventional SEM software. The authors provide a SAS macro (available for free download at [www.nd.edu/~kyuan/multilevel](http://www.nd.edu/~kyuan/multilevel)) that calculates the estimates. Consequently, level-specific fit information is provided as well parameter estimates and standard errors.

The authors contrasted their approach with simultaneous estimation in a real-data example and a Monte-Carlo simulation study. In the real data example misfit was greater for the between-groups model and the normal-theory  $\chi^2$  test statistic was severely affected by non-

normality in the data. These conclusions, the authors argue, were not made obvious by test statistics obtained with the simultaneous estimation procedure. In the simulation study, the authors generated data for a two-factor CFA model and then fit three models to the data: (1) A correctly-specified model (2) A model in which the within-group model but not the between-group model was misspecified, and (3) A model in which the between-group but not within-group model, was misspecified. The segregating approach demonstrated greater statistical power for detecting the two misspecified models as compared to the simultaneous approach for five model fit test statistics.

Yuan and Bentler (2007) note several advantages to this approach. First, the level at which misfit is occurring can be determined. Second, the calculation of commonly-used AFIs is straightforward. Third, other model diagnostics (e.g. modification indices) can be used to detect local misspecification. Fourth, misspecification at one level does not affect the evaluation of model fit at other levels. Finally, although not developed in their article, the authors suggest that the segregating approach is generalizable to an arbitrary number of data levels, given sufficient sample sizes at each. The primary disadvantage of their approach is its inability to accommodate models with random slopes. An implicit assumption of the procedure is that the variance components are orthogonal. With random slopes, this assumption is violated and the procedure may provide misleading results. Also, their procedure requires preprocessing via a SAS macro that may fail at large sample sizes (Ryu & West, 2009).

*Partially-Saturated Approach.* Yuan and Bentler (2007) suggested that for a two-level model one could saturate one level of the model to evaluate fit at the other level, and vice versa. Ryu and West (2009) formally proposed this method and developed level-specific measures of the  $\chi^2$  statistic, RMSEA, and CFI. Their method requires the analyst to fully specify one level of

the model and saturated the other level. Model estimation then proceeds simultaneously at each level, but due to the saturation at the other level(s), any misspecification reflected by the discrepancy value is specific to the level of data for which a model was specified. It is important to note that this procedure was not designed to produce parameter estimates and their corresponding standard errors, although estimates obtained using the procedure can be checked against those from simultaneous estimation for agreement.

Ryu and West (2009) also presented results from a simulation study. The study design was similar to that of Yuan and Bentler (2007) with small differences in data generating parameter values. Specifically, a two-factor CFA was generated at each level, and the three analysis models described earlier were estimated. The authors then compared mean values of the  $\chi^2$  statistic, RMSEA, and CFI over replications across the simultaneous, segregating, and partially-saturated approaches. Whereas the simultaneous procedure failed to detect misfit (underfactorization) at the between-group level, both the segregating and partially-saturated approaches successfully detected the between-group misfit. Further, parameter estimates and standard errors were comparable across methods. Values for the fit statistics were slightly higher for the segregating approach, which the authors attribute to the sample size used for estimation.

In comparing the approaches, the partially-saturated approach can be estimated using a single software package and does not require pre-processing. Also, parameter estimates obtained in the partially-saturated models can be compared to those in the overall model, providing an easy check for violation of the independence assumptions. However, as Yuan and Bentler (2007) note, misfit at the fully-specified level may bias parameters estimated in the saturated level and consequently fit information. Also, the comparability of parameter estimates and standard errors in the segregating and partially-saturated approaches have not been studied (Ryu & West, 2009).

The simulation conditions used in these studies were quite limited. Yuan and Bentler (2007) fixed the intraclass correlation coefficient (ICC) at .50 and the cluster sample size at  $J = 200$ . Also, commonly-used AFIs were not examined only in the real data example. Ryu and West (2009) also fixed the ICC at .50 and employed large overall sample sizes (the smallest overall  $N$  was 2500). Such sample sizes and ICCs are not commonly seen in applied research. Ryu (2008) notes the conditions were chosen in order to demonstrate the asymptotic utility of the level-specific procedures as opposed to maximizing generality. As such, these promising approaches to fit evaluation in MSEM require further investigation.

### *Research Questions*

The present research study will address two interrelated research questions: (1) How do fit indices perform in MSEM? (2) How comparable are the three fit evaluation methods (simultaneous estimation, segregating approach, partially-saturated approach)? The first question recognizes the additional complications inherent in multilevel data. Specifically, it is unclear how measures of model fit perform under varying levels of the ICC or sample size. Further complicating matters, there are many ways to conceptualize sample size in multilevel data (e.g., number of clusters, average within-group cluster size, total sample size, cluster/member ratio) and little is known regarding the effects of each. The second question pertains to how the various fit measures are calculated. Overall model fit provides little information as to the source or magnitude of misspecification. Given the new methods to assess level-specific fit, it is important to understand how comparable each is regarding model fit sensitivity. Previous simulation work has not explored these outcomes in the context of realistic multilevel data and is thus the focus of the current study.

### Methods

A Monte Carlo simulation study was conducted to answer the proposed research questions. Data were generated using *Mplus* version 6.0 (Muthén & Muthén, 1998-2010). Three general factors were manipulated: the intraclass correlation coefficient, sample size, and severity of model misspecification. Six levels of ICC were chosen to reflect ICCs commonly used in practice. Similarly, five sample size configurations were chosen to reflect sample sizes reported in MSEM applications. Furthermore, two types of sample size configurations were used in order to separately determine the effect of overall sample size ( $N$ ) and the cluster to within-group sample size ratio ( $J/n_j$ ). These factors result in a simulation design with  $6$  (ICC)  $\times$   $5$  (sample size)  $\times$   $2$  (configuration type) =  $60$  unique cells. Severity of misspecification was defined as a random variable as opposed to a discrete factor level and is thus not included in these calculations. The values chosen for each factor as well as the definition of misspecification severity are defined later. For each cell of these  $60$  cells,  $2000$  replications were generated. Five different analysis models that correspond to different fit assessment approaches were used to analyze the  $2000$  replications within each cell: A fully-specified multilevel model simultaneously (SIM) estimated at both levels; a partially-saturated (PS) between-group, fully-specified within-group model; a partially-saturated within-group, fully-specified between-group model; a within-group model using the segregating (SEG) approach; and a between-group model using the segregating approach. To calculate the estimated within-group and between-group covariance matrices for the SEG approach, a program was written in the R statistical language based on the SAS macro provided by Yuan and Bentler (2007).

#### *Data-Generation and Analysis Models*

A two-level confirmatory factor analysis model was used for data generation. This model is shown in Figure 1 and is similar to those used in Yuan and Bentler (2007) and Ryu and West

(2009). The variables  $y_1$ - $y_6$  are observed variables that load onto two latent constructs at each level. In order to scale the latent variables and identify the model, latent factors at both levels were fixed at 1.0. This implies that this covariance between the two factors at each level is on a correlation metric. Population parameter values will differ according to the ICC and misspecification conditions as described later.

The analysis model that was used is shown in Figure 2. In this model, the six variables are accounted for by a single latent factor at each level. Therefore, the model is correctly specified when the generating model shown in Figure 1 has a population value of 1.0 for the latent factor correlation ( $\psi_{2,1}$ ) at each level. Conceptually, this implies that the two latent factors are identical and the relationships among the six indicators can be accounted for by a single latent factor. Furthermore, the model is misspecified if the population value for the factor correlation is less than 1.0 with more misspecification arising as the value becomes smaller.

### *Study Conditions*

*ICC.* Six levels of the ICC were used in this study: .05, .10, .15, .20, .25, and .30. because ICCs rarely exceed .30 in practice (Lüdtke et al., 2008) and recent Monte Carlo studies have sampled lower ICC values accordingly (Hox et al., 2010; Lüdtke et al., 2008). Also, as the ICC value chosen in Yuan and Bentler (2007) and Ryu and West (2009) is rather high, it is of interest to investigate both level-specific fit methods under more realistic data conditions. Consequently, these lower ICC were used.

Population parameter values that correspond to these ICC levels are presented in Table 2. For example, for an ICC of .30, the within-group factor loadings were set to .8367, the within-group residual variances were set to .70, the between-group factor loadings were set to .5477, and the between-group residual variances were set to .30. The values for all ICC levels

correspond to item variances of 2.0 and communalities of .50. Because the latent factor variances are fixed to 1.0 for identification purposes, the latent variances and correlation do not contribute to the calculation of the various ICCs.

*Sample size.* Five levels of sample sizes within each of two types of sample size configurations were used in this study. As previously noted, there are several sample sizes to consider in MSEM. There is the total sample size,  $N$ , as well as the number of clusters,  $J$ , and the average number of units within each cluster,  $n_j$ . When the data are balanced,  $n_j$  also represents the number of units per cluster. For simplicity, only balanced data were generated for the present study.

For the first sample size configuration type, the overall sample size ( $N$ ) varied. Five levels of  $N$  were chosen: 500, 1000, 1500, 2000, and 2500. For each of these levels, the number of units per cluster was fixed at 20, and thus the number of clusters will increase as the total sample sizes increase. These levels of  $N$  were chosen for two reasons. First, they were chosen to reflect lower total sample sizes than seen in previous simulation research. In the Ryu and West (2009) simulation,  $N$  ranged between 2500 and 10000, and for the Yuan and Bentler (2007) simulation an  $N$  of 21100 was used with uniformly distributed within-cluster sample sizes. Ryu (2008) notes that larger sample sizes were chosen to demonstrate the asymptotic properties of their method and generate stable non-normal data conditions but that the performance of the partially-saturated approach is unknown at smaller sample sizes. Second, applications of MSEM have reported sample sizes in line with the levels chosen here (cf. Dedrick & Greenbaum, 2010; Duncan, Duncan, & Strycker, 2002; Hännqvist, Gustafsson, Muthén, & Nelson, 1994) and it is of interest to determine fit index sensitivity for typical data conditions.

The second type of sample size configuration was designed to assess the effect of varying numbers of clusters and within-group sample sizes on measures of fit at an fixed overall sample size. For  $N = 1500$ , five levels of sample sizes were included. Specifically, the number of clusters and within-cluster sample sizes were manipulated to achieve differing levels of the  $J/n_j$  ratio. Julian (2001) found that decreasing this ratio at a fixed  $N$  leads to greater inflation in the  $\chi^2$  test statistic. As AFIs are based on this statistic, it is expected that the sensitivity of each will also depend on this ratio. Thus, the following five sample size configurations ( $J/n_j$ ) were used in the current study: 50/30, 75/20, 100/15, 125/12, and 150/10. The ratio values were 1.67, 3.75, 6.67, 10.42, and 15. As a result of the simulation analyses, two separate simulations were conducted that were identical with the exception of the sample size configuration type used. Hereafter, *Simulation 1* refers to the simulation in which the overall sample size  $N$  varied and *Simulation 2* refers to the simulation in which the cluster to within-cluster sample size ratio  $J/n_j$  varied.

*Severity of Misspecification.* Quantifying the severity of model misspecification is a tricky issue (Fan & Sivo, 2005; Wu & West, 2010). Misspecification in different model parameters confounds type of misspecification with severity of misspecification (Fan & Sivo, 2005). As a consequence, severity of misspecification should be quantified as its overall discrepancy from the population model. This implies that models with different types of misspecification in terms of the number and type of parameters misspecified (as well as their magnitude) still have the same overall degree of misspecification. Meaningful comparisons can then be made as the degree of misspecification is constant across the levels of other design factors.

The Satorra and Saris (1985) method of power estimation for the  $\chi^2$  test has been used to determine the severity of model misspecification in a few studies (Enders & Finney, 2003; Fan &

Sivo, 2005, 2007; Taylor, 2008; Wu & West, 2010). This test makes use of non-central  $\chi^2$  distributions, and in particular, the noncentrality parameter ( $\lambda$ ). As previously noted, the test statistic value obtained in Equation 18 is distributed as  $\chi^2$  with degrees of freedom  $df$  under proper model misspecification. Under the condition of model misspecification, however, the value obtained in (18) is distributed as non-central  $\chi^2$  with degrees of freedom  $df$  and non-centrality parameter  $\lambda$ , such that  $\lambda = (N-1) \hat{F}_{ML}$  and represents the rightwards shift from a central  $\chi^2$  distribution. Furthermore, under the assumption that true model misspecification in such circumstances is equal to or greater than misspecification due to sampling error (MacCallum et al., 1996),  $\lambda$  represents the lack of fit of a given model in the population.

In the context of power analysis, the Satorra and Saris (1985) method requires two steps and determines the power of the  $\chi^2$  test to detect misfit in a single or several model parameters. In the first step, the estimated “population” covariance matrix is calculated from a hypothesized model with pre-specified parameter values. In the second step, a similar model is fit to the estimated covariance matrix with some model parameters fixed at a designated value, usually zero. The  $T_{ML}$  value obtained in the second step is used as an estimate of  $\lambda$ , and power is calculated by comparing a non-central  $\chi^2$  distribution defined by  $\lambda$  and the misspecified model’s  $df$  to a central  $\chi^2$  distribution with the same degrees of freedom at a given sample size. Specifically, power is defined as the area under the non-central  $\chi^2$  distribution that is to the right of the central  $\chi^2$  distribution’s critical value at a given Type I error rate  $\alpha$  (usually .05).

The same procedure can be used to quantify model misspecification in a Monte Carlo simulation study. A parameter (or set of parameters) is chosen that will be omitted or fixed in the analysis model, and a value for the parameter is set to obtain a given level of power for rejecting the model via the  $\chi^2$  test at a given sample size. However, bias resulting from fixing a parameter

at a value other than its population value may spread to other estimated parameters (Wu & West, 2010). An alternative to this procedure is the true model fixed likelihood ratio test statistic (TMFLR). The TMFLR is calculated in a manner similar to the Satorra and Saris (1985) statistic except that in the second step the other model parameters are fixed to their population value and thus preventing bias “leakage” in those parameters.

Misspecification was introduced by fitting a one-factor model to a two-factor model with varying magnitudes of factor correlation (see Figures 1 and 2). This is somewhat unusual given that most researchers introduce misspecification by fixing a non-trivial model parameter to zero. However, Fan and Sivo (2005) noted that when paths are omitted in the structural portion of the model, the SRMR is difficult to interpret as several elements in the implied covariance matrix are forced to be zero. In line with Ryu and West’s (2009) simulation design, misspecification was defined in the structural portion of the model as any departure from a one-factor model via the latent factor correlation. This circumvents the problem of zeros and provides an appropriate parameterization from which the TMFLR is calculated.

In previous simulation studies, values for the TMFLR have been calculated for different levels of severity as defined by the power to reject the null hypothesis of exact model fit via the  $\chi^2$  test. For example in Wu and West (2010) the power levels chosen were .60, .80, 1.0 and with a difference of one degree of freedom between nested models correspond to TMFLR values of 4.90, 7.85, and 38.00. In the present study, however, severity of misspecification was defined as a random variable. For each of 2000 datasets generated, a value of the between-group latent factor correlation ( $\psi_{2,1}$ ) was randomly sampled from a uniform distribution ranging from .54744 to 1.0, which correspond to power values of 1.0 to 0.0, respectively, at a sample size of 75. That is, when the latent factor correlation is set to a value of .54744 for the model in Figure 1 and the

analysis model in Figure 2 is estimated, the resulting chi-square is 38.00 and the power to reject the test of exact fit is exactly 1.0. The sample size of 75 was used to calculate the TMFLR for two reasons. First, only misspecification in the between-group model was considered. As a result of the within-group model largely accounting for the discrepancy function value (cf. Ryu & West, 2010), we could reasonably estimate the performance of various fit measures based on previous research of single-level SEM. However, it is relatively unclear how the same measures will perform when the between-group model is misspecified. For simplicity, given the number of existing conditions in the present study, the case in which both levels of the data were misspecified was not included. Thus, only between-model misspecification was considered, and thus a reasonable level-2 sample size ( $J$ ) was required. Second, 75 represents the average values of the study conditions for between-group sample sizes in *Simulation 1*, and it approximates sample sizes observed in the applied literature (Dedrick & Greenbaum, 2010; Purdy et al., 2010; Walsh et al., 2010). Interestingly, the factor correlation value chosen for the same model in Ryu and West (2009) corresponds to severe misspecification as defined by the power to reject the  $\chi^2$  test of exact fit.

### *Dependent Variables & Analyses*

Several fit indices were evaluated in the present study. Formulas for these indices are presented in Table 1. Specifically, the  $\chi^2$  statistic, RMSEA, SRMR, TLI, CFI, GFI\*, and AGFI\* were calculated for each fit evaluation method. Calculations for these indices using the SIM or SEG approaches are isomorphic to calculations in single-level SEM. Formulas for the partially-saturated  $\chi^2$  statistic, RMSEA, and CFI are provided in Ryu and West (2009).. Notable changes in calculations include the RMSEA for which sample sizes (as reflected in the denominator of the equation) differ depending on which level is being evaluated. Specifically,  $N$  is used in the

denominator for the within-group model and  $J$  is used in the denominator of the between-group model. This is also true of the GFI\*<sup>4</sup>. For the comparative fit indices TLI/CFI, an modified null model must be calculated separately. The AGFI\* is based on the GFI\* and thus does not require modification. The indices above were chosen because they reflect the most popular indices used by applied researchers (Wu & West, 2010) and they have shown good performance in simulation studies (Hu & Bentler, 1998, 1999; Taylor, 2008). Less well-known are the GFI\* and AGFI\* indices, which have been used sparingly in practice but have nevertheless shown consistently good performance in simulation studies (Hu & Bentler, 1998, 1999; Taylor, 2008).

Model convergence rates were calculated and reported. Several models resulted in inadmissible estimates (i.e., Heywood cases) and were not considered to have converged in convergence rate calculations. Furthermore, the fit index values resulting from these solutions were excluded from analyses. Two general methods were used to assess the sensitivity of fit indices. First, Loess curves based on regression smoothing procedures were generated using the R statistical package. Second, non-linear regressions were conducted using the study conditions as predictors. Details of these two methods are described in the Results section next.

## Results

### *Model Convergence*

Convergence rates for both simulations can be found in Tables 3 and 4. In *Simulation 1*, the SEG approach had the highest rates of convergence compared to the other methods. The SEG within-group model converged for all replications within each condition. The SEG between-group model demonstrated low rates of convergence when the ICC was set at .05 (range = .08 - 1.00). However, at other ICC levels the model converged for all replications at overall sample sizes greater than 1000. The SIM approach generally had poor convergence rates for ICC conditions less than .20 and sample sizes less than 1000. The rates for this approach were also

comparable to those of both PS models. All three models had similar rates across all sample sizes and ICCs, with the exception of the PS within-group model having slightly lower convergence rates as compared to the SIM and PS between-group models.

Convergence rates were larger for *Simulation 2*. Again, the SEG within-group model converged for each replication in every condition. The SEG between-group model had higher convergence rates as compared to *Simulation 1*. The model converged for all replications in ICC conditions of .10 and above. In the .05 ICC condition, the rates were relatively high (minimum = .73) and decreased as the number of clusters increased and the within-group sample size decreased. A similar pattern was found for the SIM and PS models in the .05 ICC condition. These two approaches again had comparable convergence rates across all levels of ICC and sample size. The PS within-group model had the lowest convergence rates across the study conditions, although rates were higher as compared to *Simulation 1*. The rates for this model increased with increasing numbers of clusters in the .05 ICC condition, which is in the opposite direction compared to the other methods.

### *Fit Index Sensitivity*

It was noted in the Methods section that two approaches were used to assess fit index sensitivity across the various study conditions: Loess curves and regression analyses. Results from these methods are provided next, with the Loess curves discussed first. For each method, analysis information will be presented followed by results that are broken down by specific fit measures.

### *Loess Curves*

The series of Loess curves used to assess fit index sensitivity can be found in Figures 3 through 10. Local regression fitting via weighted least squares with a span of approximately 2/3

was used for smoothing. Each figure corresponds to a different fit statistic ( $\chi^2$ ) or fit index (all others) with three different analysis models represented in separate windows. The SEG and PS within-group models were not included as they showed close fit for all indices and across all levels of misspecification. This was expected as the within-group model was properly specified. The scale changes between figures but remains constant within a given figure. Each measure of fit was plotted against model misspecification, as defined by the varying latent factor correlation  $\psi_{2,1}$ . It is worth noting that as the latent factor correlation is the actual value plotted, the figures are inverted along the x-axes. That is, deviations from the origin along the x-axis in the positive direction are indicative of *less* model misspecification. Recall that as the factor correlation approaches 1.0, the population generating model approaches the analysis model in which a single factor predicts the six indicators at each level. Three lines were plotted on each graph corresponding to different levels of the ICC (.05, .15, and .30). The overall sample size was fixed at 1500 for every curve and consisted of 75 clusters and 20 within-cluster units.

$\chi^2$ . The Loess curves for the  $\chi^2$  test statistic are shown in Figure 3. The within-group model for both the SEG and PS methods resulted in a low and constant  $\chi^2$  value across varying degrees of between-group misspecification. This is not surprising as the within-group model was properly specified for every generated dataset. The SIM model also demonstrated relatively low values of the  $\chi^2$  test statistic although increases were observed at greater levels of between-group misspecification and higher ICC values. A similar pattern was observed for the PS between-group model. Conversely, the SEG model demonstrated worse model fit at lower ICC levels.

*RMSEA*. Loess curves for the RMSEA are found in Figure 4. Low values of the RMSEA that are constant across varying levels of between-group misspecification were found for the SEG and PS within-group models. The same pattern was found for the SIM model and is similar

to findings reported by Ryu and West (2009). That is, as long as the within-group model fits the data well, the RMSEA tends to be low regardless of the severity of between-group model misspecification as it is based on the  $\chi^2$  test statistic. The  $\chi^2$  test statistic is weighted by level-specific sample sizes and thus is largely determined by the larger level-1 sample size.

Furthermore, the RMSEA for the SIM approach is calculated with the overall sample size in the denominator, thus further attenuating the value of the index even at severe levels of between-group misspecification.

The SEG between-group model demonstrated higher values of the RMSEA as compared to the previously discussed approaches. Even when the misspecification was minimal, the SEG between-group RMSEA implied poor fit. Worse fit was observed at lower ICC levels. This is again in contrast to the PS between-group model, in which lower values of the RMSEA (i.e. better fit) were observed at low ICCs. Also, it is clear from Figure 4 that at low ICC levels, the RMSEA is not sensitive to model misspecification via the PS approach. RMSEA values across the range of misspecification were all below .10 and generally clustered around .05, which is often thought of as demonstrating acceptable model fit. At increasing levels of the ICC, however, the RMSEA became increasingly sensitive to model misspecification.

*SRMR.* The within-group SRMR loess curves are found in Figure 5. Note that there is no within-group SRMR for the SEG approach but there is for the PS approach. As expected, the value is low across all levels of model misspecification for the PS within-group and SIM models as the analysis model at level-1 is properly specified. Also, the within-group SRMR is approximately zero for the PS between-group model as the within-group model is saturated.

Loess curves for the SRMR between-group model are found in Figure 6. There is no between-group SRMR for the SEG within-group model. The between-group SRMR is

approximately zero for the PS within-group model as the between part of the model was saturated. For the remaining methods, the SRMR between-group value generally increased with increasing model misspecification. Additionally, at lower levels of the ICC the between-group SRMR was higher and thus reflected worse fit for the SIM, SEG, and PS between-group models. This is the only index for the SIM and PS methods for which such a pattern was observed.

*CFI.* The CFI Loess curves are reported in Figure 7. As expected, the CFI demonstrated close fit across all levels of between-group misspecification for the SEG and PS within-group models. This same pattern also emerged for the SIM approach, again replicating the findings of Ryu and West (2009). Although sample size is not included in the calculation of the CFI as it is for the RMSEA, the  $\chi^2$  test statistic is, and consequently the index will mask any between-group misspecification even if it is severe. For the SEG between-group model, the CFI did decrease as more misspecification was introduced into the between group model. However, the effect of misspecification was less pronounced in the .05 ICC condition for which the CFI was consistently low. Similar to previously discussed fit measures, the CFI appears to always suggest a poor fitting SEG between-group model when the ICC is low. The Loess curves for the PS between-group model show a slightly different pattern. The CFI appeared to be sensitive to increasing model misspecification though less so at lower levels of the ICC. As opposed to the SEG between-group model, however, the CFI values did not appear to approach convergence for different levels of the ICC as misspecification became increasingly severe. This suggests that at severe levels of misspecification, the CFI in the PS between-group model is more sensitive to between-group misspecification than the CFI calculated in the SEG between-group model.

*TLI.* In general, the TLI behaved in a similar manner to the CFI. Values for the TLI in the SEG and PS within-group models, as well as for the SIM model, were nearly identical. Loess

curves for the TLI are presented in Figure 8. The SEG between-group model was again sensitive to varying levels of between-group misspecification. The TLI for this model was below common threshold guidelines (.90 or .95) when the ICC was set at .05 even as between-group misspecification approached zero. As with the CFI, the TLI appears to be more sensitive to between-group model misspecification when the ICC is higher for the SEG and PS between-group model. Furthermore, the PS between-group TLI did not converge to similar levels across ICCs at severe levels of misspecification as seen in the SEG between-group model.

*GFI\**. The Loess curves for the alternative GFI estimator (i.e., gamma hat) are presented in Figure 9. The *GFI\** values were near 1.0 across all levels of between-group misspecification for the SEG and PS within-group models. Interestingly, at higher levels of the ICC the *GFI\** was sensitive to increasing amounts of model misspecification for the SIM estimation approach. This is in contrast to the RMSEA and CFI as observed by Ryu and West (2009) which were not sensitive to severe levels of between-group misspecification. For the SEG between-group model, a similar pattern was observed for the *GFI\** as compared to the CFI and TLI. That is, the *GFI\** decreased monotonically when model misspecification increased, although less so for the .05 ICC condition in which the *GFI\** was at consistently unacceptable levels. For the PS between-group model, however, a different pattern emerged as compared to the CFI and TLI. At low ICC levels, the *GFI\** was not sensitive to increasing model misspecification and values suggested acceptable fit. Conversely, the CFI and TLI were somewhat sensitive to model misspecification at low ICC levels, and the values suggested poor fit across the entire range of model misspecification.

*AGFI\**. The *AGFI\** Loess curves can be found in Figure 10. The patterns for the *AGFI\** were identical to those of the *GFI\**, with the exception that *AGFI\** values were typically lower

than GFI\* values across all models, severity of misspecification, and ICC. Once again, the AGFI\* suggested excellent model fit for the SEG and PS within-group models which were properly specified and was sensitive to increasing between-group misspecification for the SIM model. The AGFI\* for the SEG between-group model was sensitive to increasing misspecification at high ICC levels. However, even at high levels of ICC and low levels of misspecification, the observed AGFI\* suggested poor model fit. Finally, the AGFI\* was insensitive to increasing misspecification for the PS between-group model at low ICC levels, but appeared to be very sensitive at higher ICC levels.

*Summary.* There are several points worth reiterating that emerged from the Loess curves. First, the fit measures indicated close model fit for the SEG and PS within-group models across all-levels of between-group misspecification. This was expected as the model for level-1 was properly specified. Second, for the SIM approach only the  $\chi^2$  test statistic, between-group SRMR, GFI\* and AGFI\* were sensitive to misspecification in the between-group model. Third, worse model fit was observed at low ICC levels for the SEG between-group model regardless of the measure used. The opposite pattern occurred for the PS between-group model, with the exception of the between-group SRMR. Finally, it appears that the ICC level interacts with the severity of misspecification across several but not all of the fit measures and fit evaluation approaches. Furthermore, greater sensitivity to between-group misspecification was observed for higher levels of the ICC when this interaction was present. These findings will be discussed in greater detail later. Next, results from the regression analyses are reported.

### *Regression Analyses*

Non-linear regressions were conducted for both *Simulations 1* and *2*. These regressions served several purposes. First, as the sample size was fixed for each Loess curve that was

generated, the fit indices' sensitivities to sample size were unclear. Second, the dependent variables and design factors were all on meaningful metrics and (for the most part) defined on an interval scale. As such, multiple regression was used instead of analysis of variance (ANOVA), the latter of which is commonly used in simulation research. Separate regressions were conducted for three of the five fit assessment methods. The SEG and PS within-group models did not have substantial variation for any of the fit indices and thus were not included in the regression analyses. Within each of the remaining three methods, a model was estimated for each fit index which was predicted by the three design factors (ICC, severity of misspecification, and sample size) as well as all possible interactions between the factors. The coefficient of determination ( $R^2$ ) and standardized regression coefficients were calculated for each model and are reported in Tables 5 (*Simulation 1*) and 6 (*Simulation 2*). The predictors in each model were uncorrelated with each other as a result of crossing the values of each level of ICC and sample size as well as the randomly drawn values for model misspecification. As such, the standardized coefficients were comparable.

$\chi^2$ . The coefficient of determination ( $R^2$ ) for the  $\chi^2$  test statistic ranged between .29 and .34 for the SIM and SEG between-group models in both simulations. For the partially-saturated between-group model,  $R^2$  was very low. This is due to the reduced amount of variance available at the between-group level that was predictable. As the ICC increased, the  $\chi^2$  also increased for the SIM model ( $\beta_s = .21$  and  $.23$ ) but decreased for the SEG between-group model ( $\beta_s = -.40$  and  $-.39$ ). This matches the patterns found in the Loess curves for the  $\chi^2$  earlier. For all models across both simulations, the  $\chi^2$  increased as the amount of between-group model misspecification increased, as expected. For the SIM approach, total sample size  $N$  accounted for more variance in the  $\chi^2$  as opposed to the  $J/n_j$  ratio ( $\beta_s = .23$  versus  $.10$ ), but the opposite was true for the SEG

between-group approach ( $\beta_s = .14$  versus  $.29$ ). Finally, in *Simulation 1* there appeared to be non-trivial interaction effects between the ICC and misspecification for the SIM model ( $\beta_s = -.14$ ) as well as between misspecification and sample size for both the SIM and SEG between-group models ( $\beta_s = -.20$  and  $-.17$ ). In *Simulation 2*, the largest interaction effect for the SIM model was between ICC and misspecification ( $\beta_s = -.17$ ) and for the SEG between-group model the largest interaction effect was between the ICC and  $J/n_j$  ratio ( $\beta_s = -.17$ ). Overall, it appears all of the study design factors predicted non-trivial variation in the  $\chi^2$  test statistic, with some interaction effects present.

*RMSEA*. The coefficient of determination ( $R^2$ ) for the RMSEA ranged between  $.18$  and  $.32$  for *Simulation 1* and between  $.29$  and  $.38$  for *Simulation 2*. Interaction effects for all methods and in both simulations were of small magnitudes, the largest resulting between the ICC and misspecification ( $\beta_s$  range =  $-.08$  and  $-.16$ ). For the SIM model and SEG between-group model, the effect of ICC was similar to those observed for the  $\chi^2$  test statistic, with the RMSEA decreasing at lower ICC levels for the SIM approach and increasing at higher ICC levels for the SEG approach. For the PS between-group model, the RMSEA decreased at lower levels of ICC in both simulations ( $\beta_s = .14$  and  $.19$ ). Across methods and both simulations, misspecification had considerable effects, ranging between  $-.34$  and  $-.49$ . It is important to remember that the larger the effects are for misspecification, the more desirable a given fit measure is as fit measures should be sensitive only to model misspecification and not extraneous data or design factors. For sample size, the RMSEA decreased with increasing  $N$  for all methods ( $\beta_s$  range =  $-.07$  to  $-.30$ ) but increased with an increasing  $J/n_j$  ratio for all methods ( $\beta_s$  range =  $.03$  to  $.10$ ). This replicates the common observation of a positive small sample bias for the RMSEA (Chen et al.,

2008) and also suggests that more clusters and less within-cluster units results in worse fit as indicated by the RMSEA.

*SRMR.* As noted in Tables 5 and 6, for the SEG between-group model, there was not a within-group SRMR. For both the SIM and PS between-group models in *Simulation 1*, exactly half of the variation in the within-group SRMR was accounted for by the set of predictors ( $R^2 = .50$ ). Furthermore, all effects on the SRMR within-group index were negligible with the exception of overall sample size which had very strong effects ( $\beta_s = -.71$ ). This finding supports those of Taylor (2008) and others who discouraged the use of the SRMR as the result of its sensitivity to sample size—as well as its lack of monotonicity—in certain instances. For *Simulation 2*, small effects were found for all predictors and interactions in the SIM model, including the  $J/n_j$  ratio. However, for the PS between-group model, there were larger effects for several of the design factors. Specifically, the SRMR within-group index decreased as the ICC and misspecification increased ( $\beta_s = -.39$  and  $-.31$ ), and increased as the  $J/n_j$  ratio increased. Finally, there appeared to be a small interaction effect between misspecification and the  $J/n_j$  ratio for the PS between-group model as well.

Results for the between-group SRMR were largely consistent across the three methods and both simulations. The  $R^2$  ranged between .38 and .47 across methods and simulations. Additionally, the SRMR between-group value decreased as the following factors increased: ICC ( $\beta_s$  range =  $-.21$  to  $-.33$ ), severity of misspecification ( $\beta_s$  range =  $-.47$  to  $-.53$ ), overall sample size ( $\beta_s$  range =  $-.31$  to  $-.37$ ), and the  $J/n_j$  ratio ( $\beta_s$  range =  $-.03$  to  $-.53$ ). Interaction effects were negligible across methods and both simulations as well. Although the between-group SRMR appears to be relatively sensitive to misspecification in the between-group model, the other design factors also appear to account for sizable amounts of its variation.

*CFI and TLI.* Results for the CFI and TLI are nearly identical and considered together here. In *Simulation 1*,  $R^2$  ranged between .15 and .33 for the CFI and between .06 and .26 for the TLI. Also for the SIM approach these indices increased (i.e. demonstrated better fit) as the ICC decreased ( $\beta_s = -.17$  and  $-.22$  for the CFI and TLI, respectively). For the SEG and PS between-group models, the CFI and TLI increased as the ICC increased ( $\beta_s$  range = .10 to .30). For all three approaches, the CFI and TLI increased as misspecification in the between-group model decreased ( $\beta_s$  CFI range = .29 to .43;  $\beta_s$  TLI range = .19 to .46) and as the overall sample size increased ( $\beta_s$  CFI range = .14 to .27;  $\beta_s$  TLI range = .10 to .16). Interaction effects for the CFI and TLI in *Simulation 1* were mostly negligible.

In *Simulation 2*,  $R^2$  ranged between .27 and .39 for both the CFI and TLI. The effects of the ICC and severity of misspecification were very similar to those observed in *Simulation 1* for both the CFI and TLI. A slight negative effect of the  $J/n_j$  ratio appeared for all three estimation methods, but the sizes of the standardized regression coefficients were smaller than for the overall sample size in *Simulation 1*. Additionally, a noticeable interaction effect is shown for ICC and misspecification. All other interaction effects had coefficients of .10 and below. In sum, the coefficients for misspecification are high for most of the conditions in the simulations for the CFI and TLI, which again is desirable. However, these indices also appear to be less sensitive to misspecification at lower ICCs and larger overall sample sizes.

*GFI\* and AGFI\*.* Similar to the CFI and TLI, the results for the GFI\* and AGFI\* were very close and are thus discussed together. The  $R^2$  values across both simulations and all three models were similar in magnitude and direction, ranging from .35 to .49. In *Simulation 1*, the GFI\* and AGFI\* increased (i.e., showed better fit) when the ICC decreased ( $\beta_s$  range = .17 to .27) in the SIM and PS between-group models. In the SEG between-group model, the GFI\* and

AGFI\* increased as the ICC simultaneously increased ( $\beta_s = .40$  and  $.17$ ). For all three models, the indices showed improved fit as both misspecification ( $\beta_s$  range =  $.35$  to  $.49$ ) and the overall sample size increased ( $\beta_s$  range =  $-.17$  to  $-.34$ ). Also, there appeared to be non-negligible interactions present for all three methods between the ICC and severity of misspecification ( $\beta_s$  range =  $.18$  to  $.23$ ) and, to a larger extent, between severity of misspecification and sample size ( $.08$  to  $.16$ ). In Simulation 2, the magnitude and direction for most of the coefficients were close to their *Simulation 1* counterparts. Of note, however, is the effect of sample size on the GFI\* and AGFI\*. In *Simulation 2*, these indices suggested worse fit as the  $J/n_j$  ratio increased, with the largest effects emerging for the SEG between-group model ( $\beta_s = -.35$  and  $-.27$ ). Also, an interaction between the ICC and misspecification produced non-trivial effects across the three different models ( $\beta_s$  range =  $.13$  to  $.20$ ). The other interaction effects were of small magnitude.

*Summary.* In summary, the regression analyses replicated the findings of the Loess curves. The additional information provided by the regressions included the effects of sample size (both  $N$  and the  $J/n_j$  ratio) and the comparative magnitudes of effects via standardized regression coefficients. Overall, the set of predictors accounted for less than half of the variation in each fit measure. As expected, all of the indices demonstrated improvement in model fit as the severity of misspecification decreased, and the magnitude of these standardized regression coefficients were the largest among all predictors in the model. The only exception to this finding was the within-group SRMR for which was determined largely by sample size. However, there was little variation observed in this index as a result of the within-group model being correctly specified, and thus other effects may be present when misspecification is introduced. The ICC and sample size configurations generally had non-trivial effects on all of the indices. For the SIM and PS between-group models, the  $\chi^2$ , RMSEA, CFI, TLI, GFI\*, and AGFI\* all

suggested better model fit as the ICC decreased. In other words, these indices were less sensitive to model misspecification as the between-group variance decreased. Conversely, the between-group SRMR actually suggested better fit as the item ICCs increased. For the SEG between-group model, all of the fit indices suggested better model fit as the ICC increased. With regards to the overall sample size  $N$ , the  $\chi^2$ , RMSEA, SRMR, CFI, and TLI suggested better model fit as  $N$  increased for the SIM, SEG between-group, and PS between-group models. However, the GFI\* and AGFI\* actually suggested worse fit as the sample size increased for the three methods. For the  $J/n_j$  ratio, all of the fit indices suggested worse fit as the ratio increased, or rather, at a fixed level-1 sample size the number of clusters increased.

### Discussion

The results of the present study suggest the ICC and sample size configuration affect measures of fit in MSEM. When the levels of a multilevel model are estimated simultaneously or via the partially-saturated approach, measures of fit appear less sensitive to detecting misspecification at lower levels of the ICC. The sole exception is the SRMR which suggests worse fit at lower levels of the ICC. Additionally, a larger overall sample size will result in improved measures of fit with the exception of the GFI\* and AGFI\* for which fit will be worse. Sample size also plays an important role in detecting misspecification at the between-group level as larger ratios of the cluster to within-cluster sample size results in worse fit for all measures considered here.

Given the reduced sensitivity to ICC as well as the inconsistency of the SRMR, the utility of MSEM at low ICC levels is called into question. Practically speaking, estimating a model with small indicator ICCs is often intractable regardless (Muthén, 1994). This was certainly the case in the present study as ICC values of .05 and often .10 resulted in low

convergence rates. However even if model convergence is successful it may not be obvious whether the between-group model fits the data well as most measures of fit appear insensitive to model misspecification.

A second and perhaps more troublesome observation is that less than half of the variation in each index was attributable to the simulation conditions which included model misspecification. Many fit indices were developed not only to measure model misspecification in a conceptually unique way, but also to reduce the effects of extraneous data and analytic factors that had confounded previously developed measures. The results presented here suggest that factors other than model misspecification may predict appreciable amounts of variation in measures of fit. This is by no means a new finding and is actually a central theme of most methodological work on fit evaluation (Chen et al., 2008; Yuan, 2005). However, the present study further explored this issue in the context of MSEM and it appears that assessing fit for multilevel data poses additional complications as compared to those previously observed for single-level SEM.

Despite such cautions in the methodological literature regarding fit indices, it is not expected that the enterprise of fit assessment will be altogether abandoned, nor should it be (Yuan, 2005). In fact, recent work has explored several promising avenues for reforming current practices. New methods have been proposed for the detection of local misspecification in SEM models using Lagrange multipliers tests or instrumental variable estimators (Bollen, Kirby, Curran, Paxton, & Chen, 2007; Saris, Satorra, & Van der Veld, 2009). Some have sought to redefine notions of model fit in terms of model complexity and offer alternative fit measures not well known in the social science literatures (Preacher, 2006). Still others have developed resampling strategies that circumvent some problems with traditional cutoff values (Millsap &

Lee, 2009). Clearly there is momentum in the field to rework the model fit paradigm within SEM. Nevertheless, current practices of assessing fit are likely to continue. As such, some preliminary guidance can be proposed based on the results of this study. Foremost, level-specific methods of fit are encouraged in order to determine the extent of misspecification at each level of data. Although the values of level-specific measures of fit will be influenced by indicator ICCs and sample size, they may nevertheless provide important information that may be missed using simultaneous estimation. Additional steps are required to implement these methods but both the segregating and partially-saturated approaches are relatively easy to implement. The segregating approach may be slightly more challenging to execute as it requires additional software, but few changes are needed to the SAS macro as provided in Yuan and Bentler (2007) or the R program provided in Appendix A.

Before level-specific fit evaluation is conducted, however, one might calculate the GFI\* and AGFI\* as these indices appear more sensitive to between-group misspecification than other commonly used indices such as the RMSEA or CFI. Furthermore, these indices should continue to receive attention in simulation experiments as they have shown good performance in simulation work despite their rare use (Taylor, 2008; Wu, West, & Taylor, in press). Finally, it should be noted that in choosing a method of level-specific fit evaluation, one must be aware of the ICC value. At high levels of the ICC such as those used in the Ryu and West (2009) and Yuan and Bentler (2007) simulations, the SEG and PS methods appear to produce similar fit index values. However results of this simulation suggest that at low ICCs levels the partially-saturated approach may underestimate misfit whereas the segregating approach is unlikely to ever suggest a well-fitting model. One possible explanation for this discrepancy relates to the observation that small unique variances can lead to inflated  $\chi^2$  values (Browne, MacCallum,

Kim, Andersen, & Glaser, 2002). When the ICC is low, the between-group item variance components are also low and by implication the unique variance components. As the SEG approach uses single-level SEM estimation this phenomenon is expected to occur. In the simulation results presented here the SRMR was the only index that did not reflect poor fit for the between-group SEG model at low ICC levels. This is because the SRMR is based on residuals as opposed to the discrepancy function value. However, it is unclear why the same phenomenon does not occur for the partially-saturated approach. This presents an interesting question for future research to consider.

### *Limitations and Strengths*

The present study had the following limitations. First, the range of factors considered was limited. Over the past 30 years, methodologists have investigated several data and analytic conditions that influence fit statistics and indices. These include, but are not limited to: sample size, the number of variables, estimation method, the type of model, cluster balance, indicator reliability, and severity of misspecification (Anderson & Gerbing, 1984; Bearden, Subhash, & Teel, 1982; Beauducel & Wittmann, 2005; Breivik, & Olsson, 2001; Chen et al., 2008; Ding, Velicer, & Harlow, 1995; Fan & Sivo, 2005, 2007; Hox & Maas, 2001; Kenny & McCoach, 2003; Sugawara & MacCallum, 1993; Taylor, 2008). A simulation investigation including all influential factors would become unwieldy, and thus the current study was restricted to a subset of these. Second, convergence rates were low for some of the study conditions, especially when the ICC or overall sample size was low. This likely reduces some precision with regard to the Loess plots, regression analyses, and parameter/standard error estimation. Third, due to the definition of misspecification severity, the ICC and sample size conditions could not be defined as random variables. Although the intervals between ICC values were equal, precision may have

nevertheless been lost in the regression analyses. Fourth, the within-group model was perfectly specified, and thus results are generalizable only to models in which negligible misspecification exists at level-1. Finally, in this simulation and nearly all others, the population (i.e., data generating) model was assumed to be correct. It has been noted that all statistical models, even those that provide a perfect fit in the population, are only useful approximations of complex phenomena (Wu & West, 2010). Therefore all models including population models used in Monte Carlo simulation experiments are wrong to some extent. There is not yet consensus on how to accommodate misspecification in the population model, although some methods have been suggested (Chun & Shapiro, 2010; Cudeck & Browne, 1992). This is an important issue that methodologists should continue to explore for applications to future simulation work.

Despite these limitations, several strengths of the study design can be noted. First, values of the conditions chosen are reflective of data to be found in applied social science areas. Previous studies of level-specific fit methods employed idealistic simulation conditions in order to validate the methods proposed (Ryu & West, 2009; Yuan & Bentler, 2007). Although important, such studies provide little practical guidance. The present study, however, offers further insight into analytic conditions that occur in practice. Second, model misspecification was included as a study condition. Including severity of misspecification in a study of fit measures allows one to determine other factors that account for variation in fit values beyond model misspecification (Browne & Cudeck, 1993). Failure to include or properly define model misspecification has limited previous findings in this area (Fan & Sivo, 2005; Taylor, 2008). However, in the present study the TFMLR was used to define severity of model misspecification and values corresponding to this definition were randomly sampled from a continuous statistical distribution. Consequently model misspecification was placed on an interpretable metric and

non-linear effects were observed via Loess plots and regression analyses. Finally, the indices chosen for investigation were both representative of applied research areas and favorable performance in previous simulation work.

#### *Future Directions*

Model fit in single-level SEM has been an active area of research for several years. With MSEM's increasing availability and tractability, it is expected that future work will also examine model fit index performance in MSEM. It will be important to determine whether findings from the single-level SEM literature replicate in MSEM. Extensions of the present study include examination of fit index sensitivity for different model types and under conditions in which both levels are misspecified. Also, the discrepancy between the segregating and partially-saturated approaches at low ICC levels poses an interesting question for future research to consider. More generally, however, it appears that momentum towards the reform of fit index and cutoff value usage will continue, and innovative methods of fit evaluation will continue to be sought and valued.

## References

- Ansari, A., Jedidi, K., & Jagpal, S. (2000). A hierarchical Bayesian approach for modeling heterogeneity in structural equation models. *Marketing Science, 19*, 328-347.
- Anderson, J. G., & Gerbing, D. W. (1984). The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika, 49*, 155-173.
- Asparouhov, T., & Muthén, B. (2007). Computationally efficient estimation of multilevel high dimensional latent variable models. *Proceedings of the 2007 Joint Statistical Meetings, Section on Statistics in Epidemiology* (pp. 2531-2535). Alexandria, VA: American Statistical Association.
- Bauer, D. J. (2003). Estimating multilevel linear models as structural equation models. *Journal of Educational and Behavioral Statistics, 28*, 135-167.
- Bearden, W. O., Sharma, S., & Teel, J. E. (1982). Sample size effects on chi square and other statistics used in evaluating causal models. *Journal of Marketing Research, 19*, 425-430.
- Beauducel, A., & Wittmann, W. W. (2005). Simulation study on fit indexes in CFA based on data with slightly distorted simple structure. *Structural Equation Modeling, 12*, 41-75.
- Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*, 238-246.
- Bentler, P. M. (1995). *EQS structural equations program manual*. Encino, CA: Multivariate Software.
- Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*, 588-606.
- Bollen, K. A. (1989). A new incremental fit index for general structural equation models. *Sociological Methods and Research, 17*, 303-316.
- Bollen, K. A., Kirby, J. B., Curran, P. J., Paxton, P. M. & Chen, F. (2007). Latent variable models under misspecification: Two-stage least squares (2SLS) and maximum likelihood (ML) estimators. *Sociological Methods and Research, 36*, 48-86.
- Breivik, E., & Olsson, U. H. (2001). Adding variables to improve fit: The effect of model size on fit assessment in LISREL. In R. Cudeck, S. Du Toit, & D. Sörbom (Eds.), *Structural equation modeling: Present and future: A Festschrift in honour of Karl Jöreskog* (pp. 169-194). Chicago: Scientific Software International.
- Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford Press.

- Browne, M. W. & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage.
- Browne, M.W., MacCallum, R. C., Kim, C., Andersen, B. L., & Glaser, R. (2002). When fit indices and residuals are incompatible. *Psychological Methods*, 7, 403-421.
- Chen, F., Curran, P. J., Bollen, K. A., Kirby, J., & Paxton, P. (2008). An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. *Sociological Methods & Research*, 36, 462-494.
- Chun, S. Y., & Shapiro, A. (2010). Construction of covariance matrices with a specified discrepancy function minimizer, with application to factor analysis. *SIAM Journal on Matrix Analysis and Applications*, 31, 1570-1583.
- Cronbach, L. J. (1976). *Research on classrooms and schools: Formulation of questions, design and analysis*. Stanford, CA: Stanford University Evaluation Consortium.
- Cudeck, R., & Browne, M. W. (1992). Constructing a covariance matrix that yields a specified minimizer and a specified minimum discrepancy value. *Psychometrika*, 57, 357-369.
- Dedrick, R. F., & Greenbaum, P. E. (2010). Multilevel confirmatory factor analysis of a scale measuring interagency collaboration of children's mental health agencies. *Journal of Emotional and Behavioral Disorders*, XX, 1-14.
- Diez-Roux, A. V. (1998). Bringing context back into epidemiology: Variables and fallacies in multilevel analyses. *American Journal of Public Health*, 88, 216-222.
- Ding, L., Velicer, W. F., & Harlow, L. L. (1995). Effects of estimation methods, number of indicators per factor, and improper solutions on structural equation modeling fit indices. *Structural Equation Modeling*, 2, 119-144.
- Duncan, S.C., Duncan, T.E., & Strycker, L.A (2002). A multilevel analysis of neighborhood context and youth alcohol problems. *Prevention Science*, 3, 125-133.
- Enders, C., & Finney, S. (2003, April). *SEM fit index criteria re-examined: An investigation of ML and robust fit indices in complex models*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.
- Fan, X. & Sivo, S. A. (2005). Sensitivity of fit indexes to misspecified structural or measurement model components: Rationale of two-index strategy revisited. *Structural Equation Modeling*, 12, 343-367.
- Fan, X., & Sivo, S. A. (2007). Sensitivity of fit indices to model misspecification and model types. *Multivariate Behavioral Research*, 42, 509-529.

- Fan, X., Thompson, B., & Wang, L. (1999). The effects of sample size, estimation methods, and model specification on SEM fit indices. *Structural Equation Modeling*, 6, 56-83.
- Goldstein, H. (1987). *Multilevel models in educational and social research*. London, England: Griffin.
- Goldstein, H. (1995). *Multilevel statistical models*. New York, NY: Halsted.
- Goldstein, H., & McDonald, R. P. (1988). A general model for the analysis of multilevel data. *Psychometrika*, 53, 455-467.
- Härnqvist, K. (1978). Primary mental abilities of collective and individual levels. *Journal of Educational Psychology*, 70, 706-716.
- Härnqvist, K., Gustafsson, J. E., Muthén, B.O., & Nelson, G. (1994). Hierarchical models of ability at class and individual levels. *Intelligence*, 18, 165-187.
- Hox, J. J. (1993). Factor analysis of multilevel data: Gauging the Muthén model. In J. H. L. Oud & R.A.W. van Blokland-Vogeleang (Eds.), *Advances in longitudinal and multivariate analysis in the behavioral sciences* (pp. 141-156). Nijmegen: ITS.
- Hox, J. J. (1998). Multilevel modeling: When and why. In I. Balderjahn, R. Mathar & M. Schader (Eds.), *Classification, data analysis, and data highways* (pp.147-154). New York: Springer Verlag.
- Hox, J. J., & Maas, C. J. M. (2001). The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples. *Structural Equation Modeling*, 8, 157-174.
- Hox, J. J., & Maas, C. J. M. (2004). Multilevel structural equation models: The limited information approach and the multivariate multilevel approach. In K. van Montfort, J. Oud, & A. Satorra (Eds.), *Recent developments on structural equation models: Theory and applications* (pp. 135-149). London: Kluwer.
- Hox, J. J., Maas, C. J. M., & Brinkhuis, M. J. S. (2010). The effect of estimation method and sample size in multilevel structural equation modeling. *Statistica Neerlandica*, 64, 157-170.
- Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, 3, 424-453.
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6, 1-55.
- Huber, P. J. (1967). The behavior of maximum likelihood estimates under non-standard conditions. *Proceedings of the 5th Berkeley symposium on Mathematical Statistics and Probability*, 1, 221-233.

- Jedidi, K., & Ansari, A. (2001). Bayesian structural equation models for multilevel data. In G. A. Marcoulides & R. E. Schumacker (Eds.), *New developments and techniques in structural equation modeling* (pp. 129-157). New York, NY: Erlbaum.
- Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, *34*, 183-202.
- Jöreskog, K. G., & Sörbom, D. (1981). *LISREL V: Analysis of linear structural relationships by maximum likelihood and least squares methods*. Chicago: International Educational Services.
- Julian, M. W. (2001). The consequences of ignoring multilevel data structures in nonhierarchical covariance modeling. *Structural Equation Modeling*, *8*, 325-352.
- Kaplan, D., Kim, J.-S., & Kim, S.-Y. (2009). Multilevel latent variable modeling: Current research and recent developments. In R. Millsap & A. Maydeu-Olivares (Eds.), *Sage handbook of quantitative methods in psychology* (pp. 592-612). Thousand Oaks, CA: Sage.
- Kenny, D. A., & McCoach, D. B. (2003). Effect of the number of variables on measures of fit in structural equation modeling. *Structural Equation Modeling*, *10*, 333-351.
- Kline, R. B. (2011). Convergence of structural equation modeling and multilevel modeling. In M. Williams (Ed.), *Handbook of Methodological Innovation* (pp. XXX-XXX). Thousand Oaks, CA: Sage.
- La Du, T. J., & Tanaka, J. S. (1989). Influence of sample size, estimation method, and model specification on goodness-of-fit assessments in structural equation models. *Journal of Applied Psychology*, *74*, 625-635.
- Lee, S. Y. (1990). Multilevel analysis of structural equation models. *Biometrika*, *77*, 763-772.
- Lee, S.-Y., & Poon, W. Y. (1998). Analysis of two-level structural equation models via EM type algorithms. *Statistica Sinica*, *8*, 749-766.
- Leite, W. L., & Stapleton, L. M. (2006). Sensitivity of fit indices to detect misspecifications of growth shape in latent growth modeling. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.
- Liang, J., & Bentler, P. M. (2004). An EM algorithm for fitting two-level structural equation models. *Psychometrika*, *69*, 101-122.
- Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B.O. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods*, *13*, 203-229.

- MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1*, 130-149.
- Maiti, S. S., & Mukherjee, B. N. (1991). Two new goodness-of-fit indices for covariance matrices with linear structures. *British Journal of Mathematical and Statistical Psychology, 44*, 153-180.
- Marsh, H. W., Balla, J. R., & Hau, K. T. (1996). An evaluation of incremental fit indices: A clarification of mathematical and empirical properties. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 315-353). Erlbaum: Mahwah, NJ.
- Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin, 103*, 391-410.
- Marsh, H. W., Hau, K. T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling, 11*, 320-341.
- McDonald, R. P. (1993). A general model for two level data with responses missing at random. *Psychometrika, 58*, 75-585.
- McDonald, R. P. (1994) The bi-level reticular action model for path analysis with latent variables. *Sociological Methods and Research, 22*, 399-413.
- McDonald, R. P., & Goldstein, H. (1989). Balanced versus unbalanced designs for linear structural relations in two-level data. *British Journal of Mathematical, 42*, 215-232.
- Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equation modeling. *Psychological Methods, 10*, 259-284.
- Millsap, R.E. & Lee, S. (July, 2009). Approximate fit in SEM without cutpoints. Presented at the Annual meeting of the Psychometric Society, Cambridge University, Cambridge, England.
- Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika, 54*, 557-585.
- Muthén, B. O. (1990). *Mean and covariance structure analysis of hierarchical data* (UCLA Statistics Series (No. 62). Los Angeles: University of California.
- Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods and Research, 22*, 376-398.
- Muthén, B. O., & Asparouhov, T. (2008). Growth mixture modeling: Analysis with non-Gaussian random effects. In G. Fitzmaurice, M. Davidian, G. Verbeke, & G. Molenberghs (Eds.), *Longitudinal data analysis* (pp. 143-165). Boca Raton, FL: Chapman & Hall/CRC.

- Muthén, B. O., & Asparouhov, T. (2011). Beyond multilevel regression modeling: Multilevel analysis in a general latent variable framework. In J. Hox & J.K. Roberts (eds), *Handbook of Advanced Multilevel Analysis* (pp. 15-40). New York: Taylor and Francis.
- Muthén, L. K., & Muthén, B. O. (1998-2010). *Mplus user's guide: Statistical analysis with latent variables* (6th ed.). Los Angeles, CA: Muthén & Muthén.
- Nevitt, J. and G. R. Hancock. (2000). Improving the root mean square error of approximation for nonnormal conditions in structural equation modeling. *Journal of Experimental Education*, 68, 251-268.
- Olsson, U. H., Foss, T., Troye, S. V., & Howell, R. D. (2000). The performance of ML, GLS, and WLS estimation in structural equation modeling under conditions of misspecification and nonnormality. *Structural Equation Modeling*, 7, 557-595.
- Olsson, U. H., Troye, S. V., & Howell, R. D. (1999). Theoretic fit and empirical fit: The performance of maximum likelihood versus generalized least squares in structural equation models. *Multivariate Behavioral Research*, 34, 31-59.
- Preacher, K. J. (2006). Quantifying parsimony in structural equation modeling. *Multivariate Behavioral Research*, 41, 227–259. doi:10.1207/s15327906mbr4103\_1.
- Preacher, K. J. (in press). Multilevel SEM strategies for evaluating mediation in three-level data. *Multivariate Behavioral Research*.
- Preacher, K. J., Zyphur, M. J., & Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods*, 15, 209-233.
- Purdy, N., Laschinger, H. K. S., Finegan, J., Kerr, M., & Olivera, F. (2010). Effects of work environments on nurse and patient outcomes. *Journal of Nursing Management*, 18, 901-913
- Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika*, 69, 167-190.
- Raudenbush, S. W. (1995). Maximum likelihood estimation for unbalanced multilevel covariance structure models via the EM algorithm. *British Journal of Mathematical and Statistical Psychology*, 48, 359-370.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Newbury Park, CA: Sage.
- Raudenbush, S. W., Rowan, B., & Kang, S. J. (1991). A multilevel, multivariate model of studying school climate with estimation via the EM algorithm and application to U.S. high school data. *Journal of Educational Statistics*, 16, 296-330.

- Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, 15, 351-357.
- Ryu, E. (2008). *Evaluation of model fit in multilevel structural equation modeling: Level-specific model fit evaluation and the robustness to nonnormality*. (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses database. (AAT 3327254)
- Ryu, E., & West, S. G. (2009). Level-specific evaluation of model fit in multilevel structural equation modeling. *Structural Equation Modeling*, 16, 583-601.
- Saris, W.E., Satorra, A., & Van der Veld, W. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling*, 16, 561-582.
- Satorra, A., & Saris, W. E. (1985). Power of the likelihood ratio test in covariance structure analysis. *Psychometrika*, 50, 83-90.
- Schmidt, W. H. (1969). *Covariance structure analysis of the multivariate random effects model* (Unpublished doctoral dissertation). University of Chicago.
- Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B*, 13, 238-241.
- Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage.
- Steiger, J. H. (1989). *EZPATH: A supplementary module for SYSTAT and SYGRAPH*. Evanston, IL: SYSTAT.
- Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, 25, 173-180.
- Steiger, J. H. & Lind, J. C. (1980). *Statistically-based tests for the number of common factors*. Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA.
- Sugawara, H. M., & MacCallum, R. C. (1993). Effect of estimation method on incremental fit indexes for covariance structure models. *Applied Psychological Measurement*, 17, 365-377.
- Sun, J. (2005). Assessing goodness of fit in confirmatory factor analysis. *Measurement and Evaluation in Counseling and Development*, 37, 240-256.
- Tanaka, J. S. (1993). Multifaceted conceptions of fit. In K. A. Bollen & J. S. Long (Eds.) *Testing structural equation models* (pp. 10-39). Newbury Park, CA: Sage.
- Taylor, A. B. (2008). *Two new methods of studying the performance of SEM fit indices*. (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses database. (AAT 3318439)

- Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*, 1-10.
- Walsh, B. M., Matthews, R. A., Tuller, M. D., Parks, K. M., & McDonald, D. P. (2010). A multilevel model of the effects of equal opportunity climate on job satisfaction in the military. *Journal of Occupational Health Psychology*, *15*, 191-207.
- White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, *50*, 1-25.
- Wu, W., & West, S. G. (2010). Sensitivity of SEM fit indices to misspecifications in growth curve models: A simulation study. *Multivariate Behavioral Research*, *45*, 420-452.
- Wu, W., West, S. G., & Taylor, A. B. (2009). Evaluating model fit for growth curve models: Integration of fit indices from SEM and MLM frameworks. *Psychological Methods*, *14*, 183-201.
- West, S. G., Wu, W., & Taylor, A.B. (in press). Model fit and model selection in structural equation modeling. To appear in R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. XX-XX). New York, Guilford Press.
- Yuan, K.-H. (2005). Fit indices versus test statistics. *Multivariate Behavioral Research*, *40*, 115-148.
- Yuan, K.-H. & Bentler, P.M. (1998). Robust mean and covariance structure analysis. *British Journal of Mathematical and Statistical Psychology*, *51*, 63-88.
- Yuan, K.-H., & Bentler, P. M. (2007). Multilevel covariance structure analysis by fitting multiple single-level models. *Sociological Methodology*, *37*, 53-82.
- Yuan, K.-H., & Hayashi, K. (2005). On Muthén's maximum likelihood for two-level covariance structure models. *Psychometrika*, *70*, 147-167.

## Footnotes

<sup>1</sup> Some three-level models are currently estimable in Mplus. Multilevel latent growth curves are one example in which repeated measures are nested within an individual, who is further nested within a higher level unit (e.g., school). These models capitalize on the SEM specification of latent growth curves in which the random effects for the growth parameters are specified as latent variables. This specification allows information pertaining to repeated measures and individuals to be captured in a single-level SEM, or in the case of a multilevel latent growth curve, at the within-group level (Muthén & Asparouhov, 2011). Also, some 3-level mediation analyses are able to be specified using unconventional model specifications (Preacher, in press).

<sup>2</sup> The notation used in this section follows that of Preacher et al. (2010). Muthén & Asparouhov (2008) describe their model within the context of latent growth curve analysis to draw parallels between parameters in the mixed linear model and their new general MSEM model.

<sup>3</sup> Exceptions to this notion are the tests based on the RMSEA proposed by MacCallum, Sugawara, & Browne (1996). Specifically, three tests of model fit were proposed based on differing null hypotheses. However, it has recently been argued that such tests are not suitable for all modeling situations as the RMSEA and its associated confidence interval are influenced not only by sample size but aspects of the model itself (Chen et al., 2008).

<sup>4</sup> Algebraic derivations for the partially-saturated RMSEA and CFI are found in Ryu (2008). The derivation of the partially-saturated GFI\* follows that of the RMSEA and is not shown here.

Table 1  
*The Chi-square Test Statistic and Other Approximate Fit Indices*

Source	Name	Formula
Jöreskog (1969)	$\chi^2$	$(N - 1)F_{ML}$
Steiger & Lind (1980)	RMSEA	$\sqrt{\frac{\max(\chi^2 - df, 0)}{df(N - 1)}}$
Bentler (1995)	SRMR	$[p^{*-1} (\mathbf{e}'\mathbf{W}_s\mathbf{e})]^{1/2}$
Maiti & Mukherjee (1991)	GFI*	$\frac{p}{p + 2 \left( \frac{\chi^2 - df}{N - 1} \right)}$
Maiti & Mukherjee (1991)	AGFI*	$1 - \frac{p^*}{df} (1 - GFI^*)$
Bentler (1990)	CFI	$\frac{\max(\chi_0^2 - df_0, 0) - \max(\chi_H^2 - df_H, 0)}{\max(\chi_0^2 - df_0, 0)}$
Tucker & Lewis (1973)	TLI	$\frac{(\chi_0^2/df_0) - (\chi_H^2/df_H)}{(\chi_0^2/df_0)}$

*Note.*  $\chi^2$  = chi-square; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual; GFI\* = revised goodness of fit index; revised adjusted goodness of fit index; CFI = comparative fit index; TLI = Tucker-Lewis index;  $p^*$  = number of non-duplicated elements in the covariance matrix;  $\mathbf{e}$  = vector of residuals from comparing the sample and model-implied covariance matrices;  $\mathbf{W}_s$  = diagonal weight matrix used to standardize elements in sample covariance matrix;  $p$  = number of variables;  $df$  = degrees of freedom;  $\chi_0^2$  = chi-square value for null model;  $df_0$  = degrees of freedom for null model;  $\chi_H^2$  = chi-square value for hypothesized model;  $df_H$  = degrees of freedom for hypothesized model;

Table 2  
*Population Parameter Values for Data Generation*

ICC	$\lambda_W$	$\theta_W$	$\lambda_B$	$\theta_B$
.05	.9745	.95	.2236	.05
.10	.9487	.90	.3162	.10
.15	.9220	.85	.3873	.15
.20	.8944	.80	.4472	.20
.25	.8660	.75	.5000	.25
.30	.8367	.70	.5477	.30

<sup>a</sup> Power calculated based on a within-group sample size of 1500

<sup>b</sup> Power calculated based on a within-group sample size of 75

*Note.* ICC = Intraclass Correlation Coefficient;  $\lambda_W$  = Within-group factor loadings;  $\theta_W$  = Within-group residual variances;  $\lambda_B$  = Between-group factor loadings;  $\theta_B$  = Between-group residual variances. All item total variances equal 2.0. Misspecification was introduced via random sampling of the latent factor correlation ( $\psi_{2,1}$ ) for values between .54744 and 1.0, which corresponds to power levels of 1.0 and 0.0, respectively, for rejecting the likelihood ratio ( $\chi^2$ ) test.

Table 3  
*Proportion of Successfully Converged Models for Simulation 1*

ICC	N	S.E.	Seg_W	Seg_B	PS_W	PS_B
.05	500	.35	1.00	.08	.35	.33
	1000	.61	1.00	.62	.57	.61
	1500	.77	1.00	.91	.70	.77
	2000	.87	1.00	.99	.79	.87
	2500	.92	1.00	1.00	.86	.91
.10	500	.59	1.00	.65	.53	.58
	1000	.89	1.00	1.00	.75	.89
	1500	.96	1.00	1.00	.89	.97
	2000	.99	1.00	1.00	.94	.99
	2500	.99	1.00	1.00	.97	1.00
.15	500	.75	1.00	.92	.61	.75
	1000	.96	1.00	1.00	.86	.96
	1500	.99	1.00	1.00	.94	.99
	2000	1.00	1.00	1.00	.98	1.00
	2500	1.00	1.00	1.00	.99	1.00
.20	500	.83	1.00	.99	.65	.82
	1000	.99	1.00	1.00	.89	.98
	1500	1.00	1.00	1.00	.95	1.00
	2000	1.00	1.00	1.00	.98	1.00
	2500	1.00	1.00	1.00	.99	1.00
.25	500	.88	1.00	1.00	.70	.87
	1000	.99	1.00	1.00	.91	.99
	1500	1.00	1.00	1.00	.97	1.00
	2000	1.00	1.00	1.00	.99	1.00
	2500	1.00	1.00	1.00	1.00	1.00
.30	500	.91	1.00	1.00	.71	.90
	1000	.99	1.00	1.00	.92	.99
	1500	1.00	1.00	1.00	.97	1.00
	2000	1.00	1.00	1.00	.99	1.00
	2500	1.00	1.00	1.00	1.00	1.00

*Note.* All clusters are of size  $n_j = 20$ . N = Total sample size; SE = Simultaneous estimation; Seg\_W = Segregating approach within-group model; Seg\_B = Segregating approach between-group model; PS\_W = Partially-saturated within-group model; PS\_B = Partially-saturated between-group model.

Table 4  
*Proportion of Successfully Converged Models for Simulation 2*

ICC	J	S.E.	Seg_W	Seg_B	PS_W	PS_B
.05	30	.79	1.00	.95	.56	.80
	60	.78	1.00	.90	.70	.77
	90	.75	1.00	.90	.80	.76
	120	.72	1.00	.83	.86	.72
	150	.68	1.00	.73	.89	.69
.10	30	.96	1.00	1.00	.76	.97
	60	.96	1.00	1.00	.88	.95
	90	.97	1.00	1.00	.95	.96
	120	.96	1.00	1.00	.97	.97
	150	.96	1.00	1.00	.98	.95
.15	30	.99	1.00	1.00	.85	.99
	60	.99	1.00	1.00	.94	.99
	90	.99	1.00	1.00	.97	.99
	120	.99	1.00	1.00	.99	.99
	150	1.00	1.00	1.00	.99	.99
.20	30	.99	1.00	1.00	.89	.99
	60	1.00	1.00	1.00	.96	1.00
	90	1.00	1.00	1.00	.99	1.00
	120	1.00	1.00	1.00	.99	1.00
	150	1.00	1.00	1.00	1.00	1.00
.25	30	1.00	1.00	1.00	.90	1.00
	60	1.00	1.00	1.00	.97	1.00
	90	1.00	1.00	1.00	.99	1.00
	120	1.00	1.00	1.00	1.00	1.00
	150	1.00	1.00	1.00	1.00	1.00
.30	30	1.00	1.00	1.00	.93	1.00
	60	1.00	1.00	1.00	.98	1.00
	90	1.00	1.00	1.00	.99	1.00
	120	1.00	1.00	1.00	1.00	1.00
	150	1.00	1.00	1.00	1.00	1.00

*Note.* The total sample size is 1500 for all conditions. J = Number of clusters; SE = Simultaneous estimation; Seg\_W = Segregating approach within-group model; Seg\_B = Segregating approach between-group model; PS\_W = Partially-saturated within-group model; PS\_B = Partially-saturated between-group model.

Table 5  
*Standardized Regression Coefficients for Simulation 1*

Predictors	$\chi^2$	RMSEA	SRMR <sub>W</sub>	SRMR <sub>B</sub>	CFI	TLI	GFI*	AGFI*
Simultaneous								
$R^2$	.33	.24	.50	.38	.15	.14	.38	.40
ICC	.21	.22	-.00	-.29	-.18	-.17	-.22	-.27
MISS	-.40	-.40	-.01	-.47	.29	.29	.44	.39
SS	.23	-.07	-.71	-.32	.14	.14	-.25	-.30
ICC*MISS	-.14	-.11	.02	.03	.11	.11	.14	.16
ICC*SS	.05	-.03	-.03	.03	.05	.05	-.04	-.04
MISS*SS	-.20	-.10	-.01	-.07	.04	.04	.21	.23
ICC*MISS*SS	-.07	-.03	-.00	.01	.01	.01	.06	.07
Segregating: Between-Group								
$R^2$	.31	.32	--	.45	.33	.06	.40	.35
ICC	-.40	-.37	--	-.21	.30	.14	.40	.17
MISS	-.31	-.34	--	-.53	.43	.19	.40	.35
SS	.14	-.30	--	-.37	.27	.10	-.17	-.34
ICC*MISS	-.03	-.08	--	-.05	.04	.02	.08	.13
ICC*SS	-.02	.08	--	-.01	-.05	-.03	-.01	.15
MISS*SS	-.17	-.09	--	-.09	.07	.04	.18	.20
ICC*MISS*SS	-.00	-.01	--	.00	-.00	-.00	.03	.04
Partially-Saturated: Between-Group								
$R^2$	.00	.18	.50	.38	.22	.26	.36	.36
ICC	.01	.14	-.00	-.28	.10	.12	-.17	-.22
MISS	-.04	-.36	.00	-.47	.43	.46	.49	.42
SS	.01	-.13	-.71	-.31	.17	.16	-.16	-.23
ICC*MISS	-.01	-.08	-.00	.03	.02	.05	.12	.15
ICC*SS	.00	-.01	-.02	.04	-.01	-.01	-.05	-.04
MISS*SS	-.02	-.08	-.00	-.07	.07	.11	.21	.23
ICC*MISS*SS	.00	-.01	.00	.02	-.02	-.03	.04	.06

*Note.* ICC = Intraclass correlation coefficient; MISS = Misspecification; SS = Sample size (N)

Table 6  
*Standardized Regression Coefficients for Simulation 2*

Predictors	$\chi^2$	RMSEA	SRMR <sub>W</sub>	SRMR <sub>B</sub>	CFI	TLI	GFI*	AGFI*
Simultaneous								
$R^2$	.29	.37	.02	.38	.27	.27	.39	.38
ICC	.23	.26	-.04	-.33	-.22	-.22	-.27	-.27
MISS	-.41	-.49	-.06	-.53	.39	.39	.49	.49
SS	.10	.10	.10	-.13	-.11	-.11	-.11	-.11
ICC*MISS	-.17	-.16	.04	.04	.18	.18	.17	.17
ICC*SS	.08	.07	-.03	-.08	-.08	-.08	-.08	-.08
MISS*SS	-.09	-.08	-.04	-.06	.10	.10	.09	.09
ICC*MISS*SS	-.06	-.05	.02	.02	.07	.07	.05	.05
Segregating: Between-Group								
$R^2$	.34	.38	--	.47	.39	.39	.49	.34
ICC	-.39	-.44	--	-.25	.35	.35	.40	.23
MISS	-.26	-.41	--	-.62	.52	.52	.44	.42
SS	.29	.03	--	-.12	-.00	-.01	-.35	-.27
ICC*MISS	-.03	-.11	--	-.07	.05	.05	.13	.20
ICC*SS	-.17	-.03	--	-.03	.02	.02	.05	-.05
MISS*SS	-.10	-.01	--	-.02	.02	.01	.07	.00
ICC*MISS*SS	-.01	-.01	--	-.01	.00	.00	.04	.04
Partially-Saturated: Between-Group								
$R^2$	.01	.29	.46	.38	.28	.31	.41	.40
ICC	.04	.19	-.39	-.32	.09	.11	-.24	-.31
MISS	-.09	-.47	-.31	-.52	.51	.53	.55	.50
SS	.02	.10	.42	-.03	-.06	-.06	-.10	-.08
ICC*MISS	-.03	-.13	.10	.04	.05	.08	.16	.18
ICC*SS	.02	.02	-.07	-.07	.01	.01	-.06	-.06
MISS*SS	-.03	.01	-.15	-.06	.06	.05	.09	.06
ICC*MISS*SS	-.02	-.02	.04	.03	.01	.02	.04	.03

*Note.* ICC = Intraclass correlation coefficient; MISS = Misspecification; SS = Sample size (N)

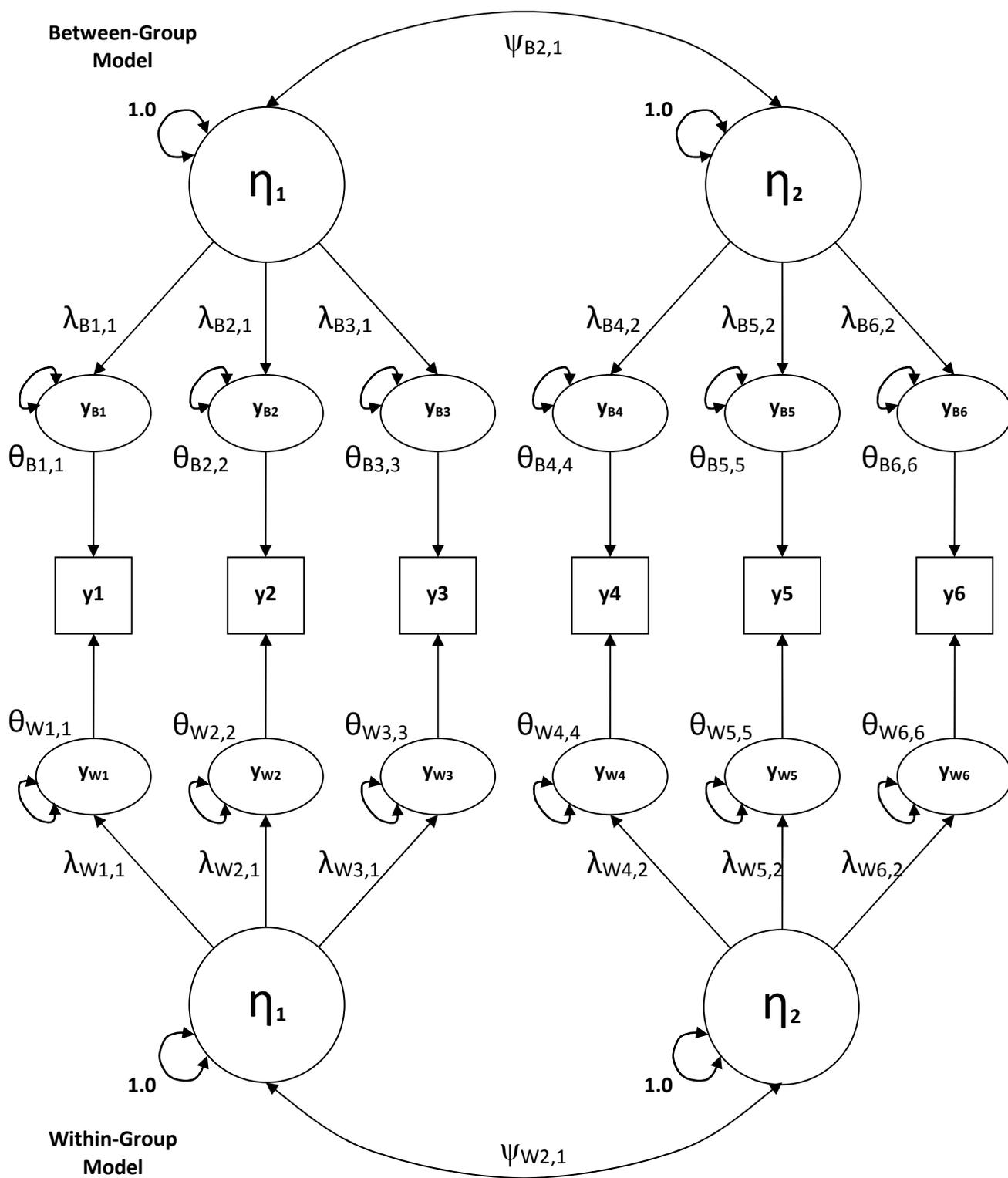


Figure 1. Population model used for data generation.

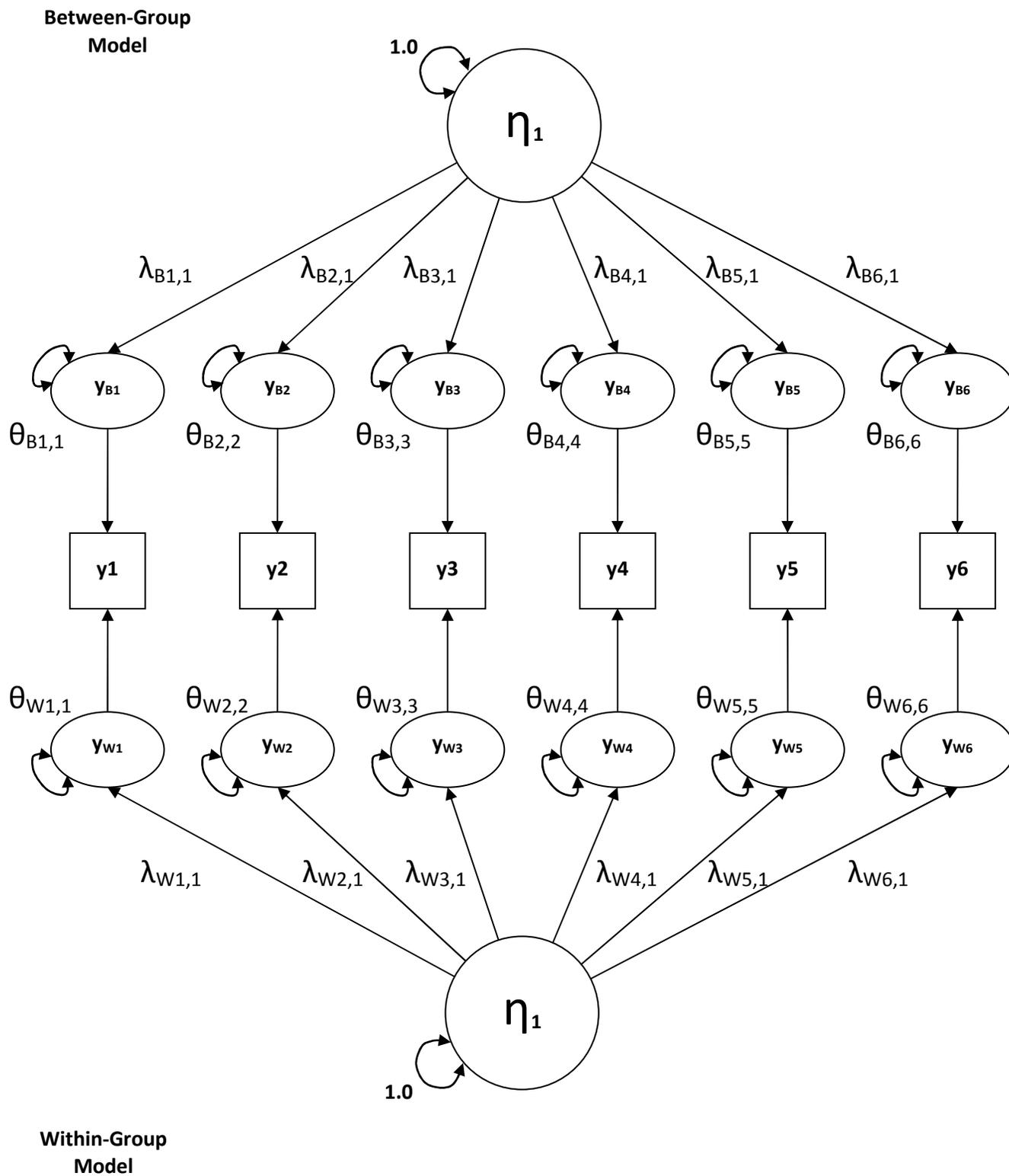


Figure 2. Analysis model.

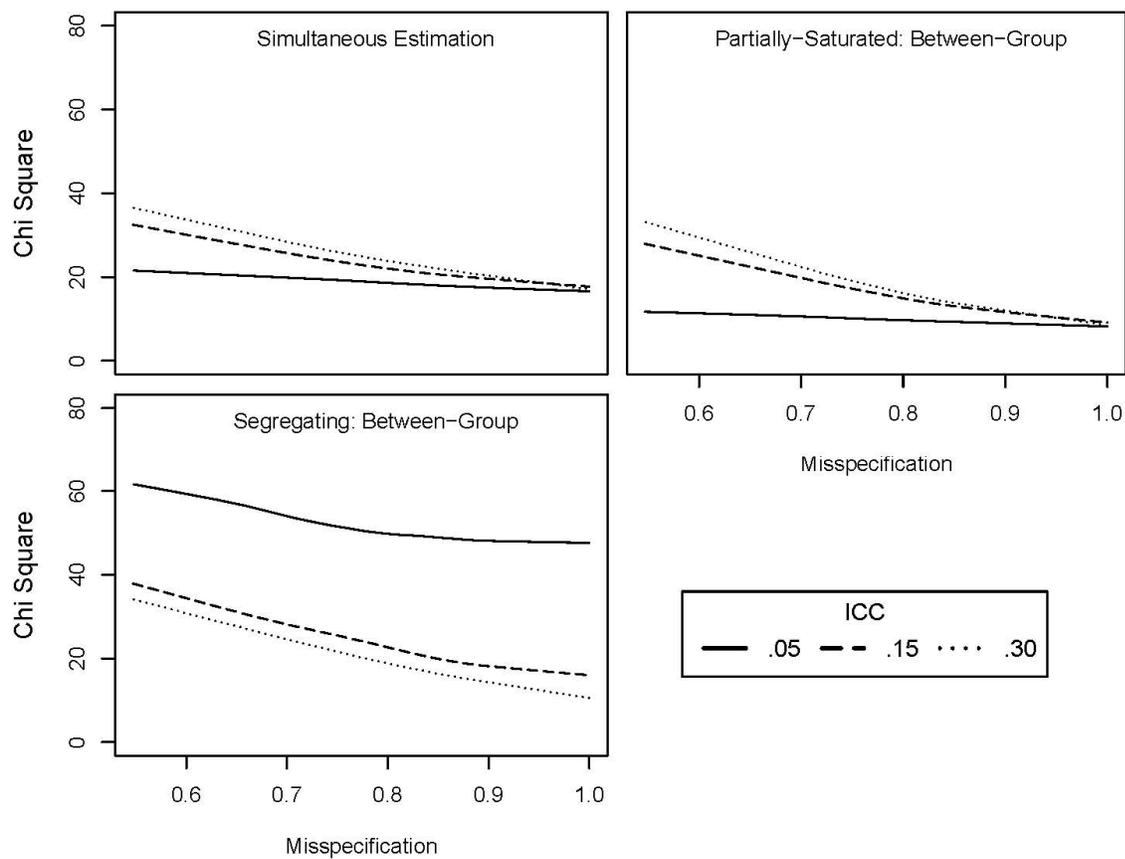


Figure 3. Loess Curves for  $\chi^2$ .

Note: Misspecification is defined on the x-axis by the value of the latent correlation  $\psi_{2,1}$ . A higher correlation value corresponds to less model misspecification.

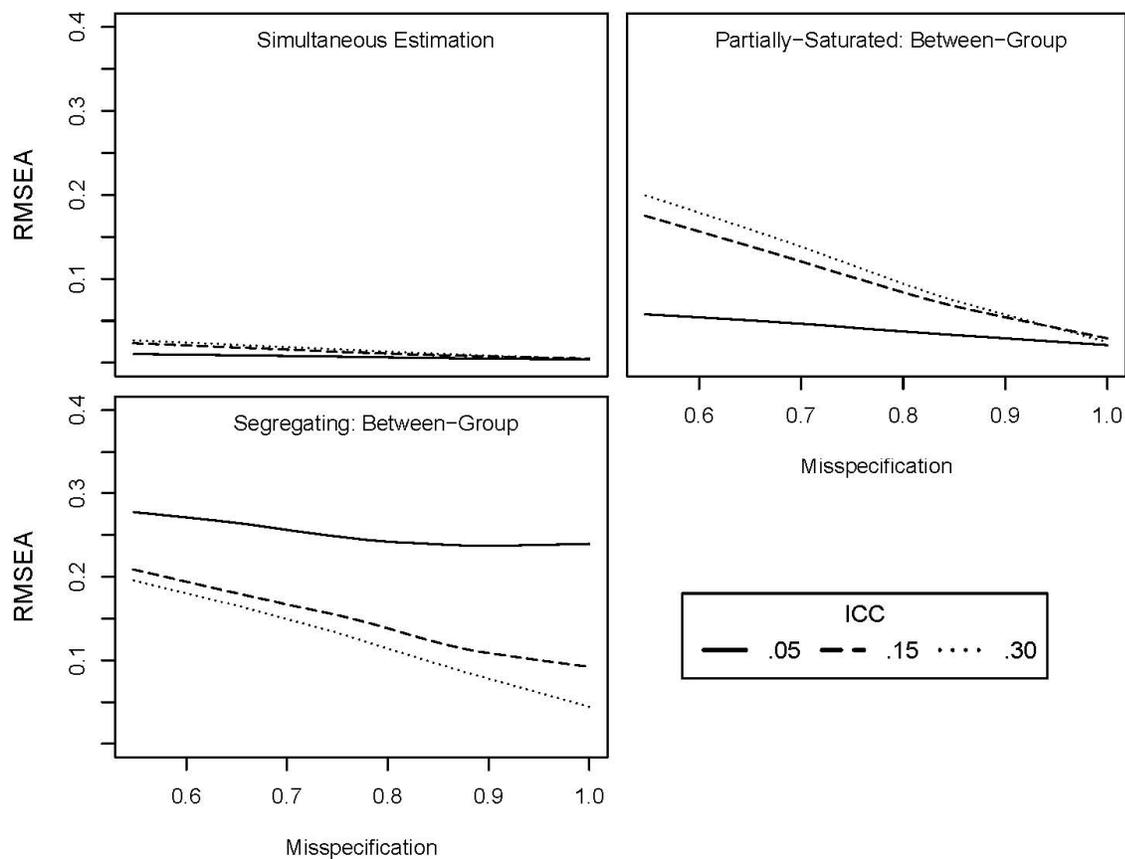


Figure 4. Loess Curves for RMSEA.

Note: Misspecification is defined on the x-axis by the value of the latent correlation  $\psi_{2,1}$ . A higher correlation value corresponds to less model misspecification.

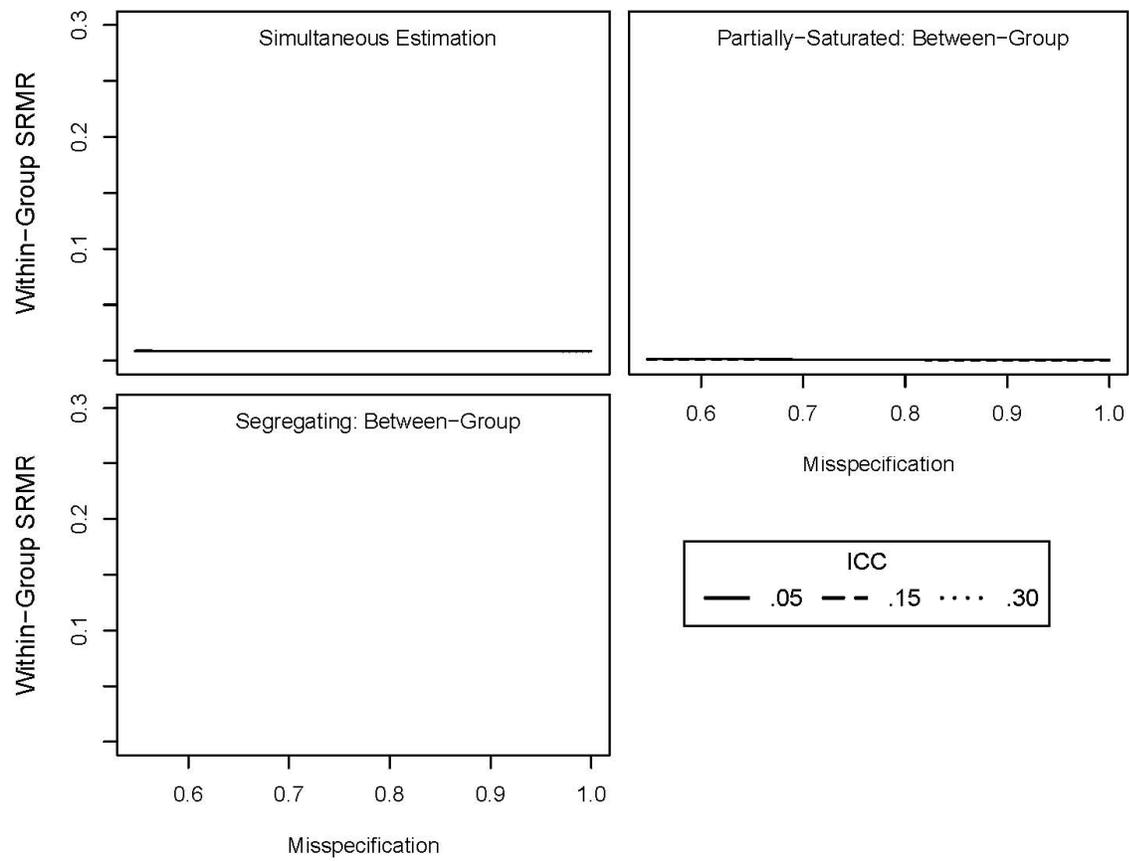


Figure 5. Loess Curves for Within-group SRMR.

Note: Misspecification is defined on the x-axis by the value of the latent correlation  $\psi_{2,1}$ . A higher correlation value corresponds to less model misspecification. The Within-group SRMR is not calculated for the segregating Between-group model.

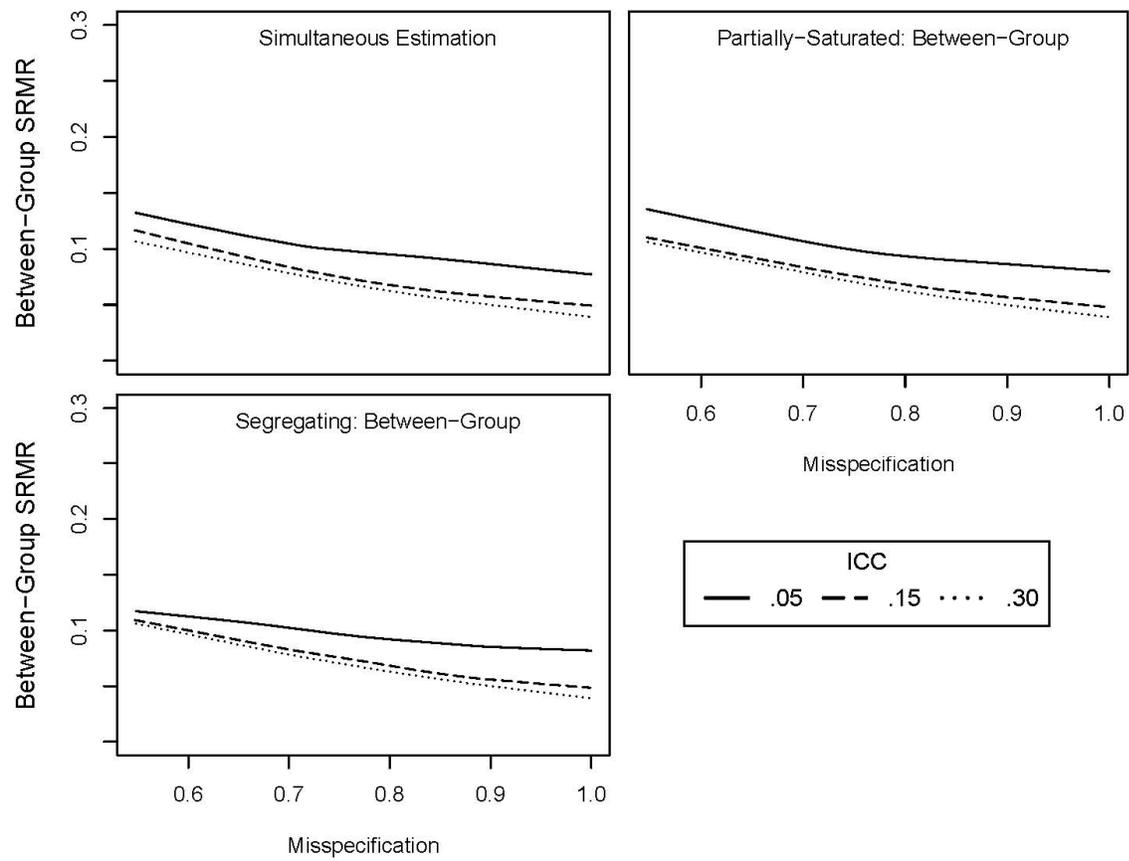


Figure 6. Loess Curves for Between-group SRMR.

Note: Misspecification is defined on the x-axis by the value of the latent correlation  $\psi_{2,1}$ . A higher correlation value corresponds to less model misspecification. The Between-group SRMR is not calculated for the segregating within-group model.

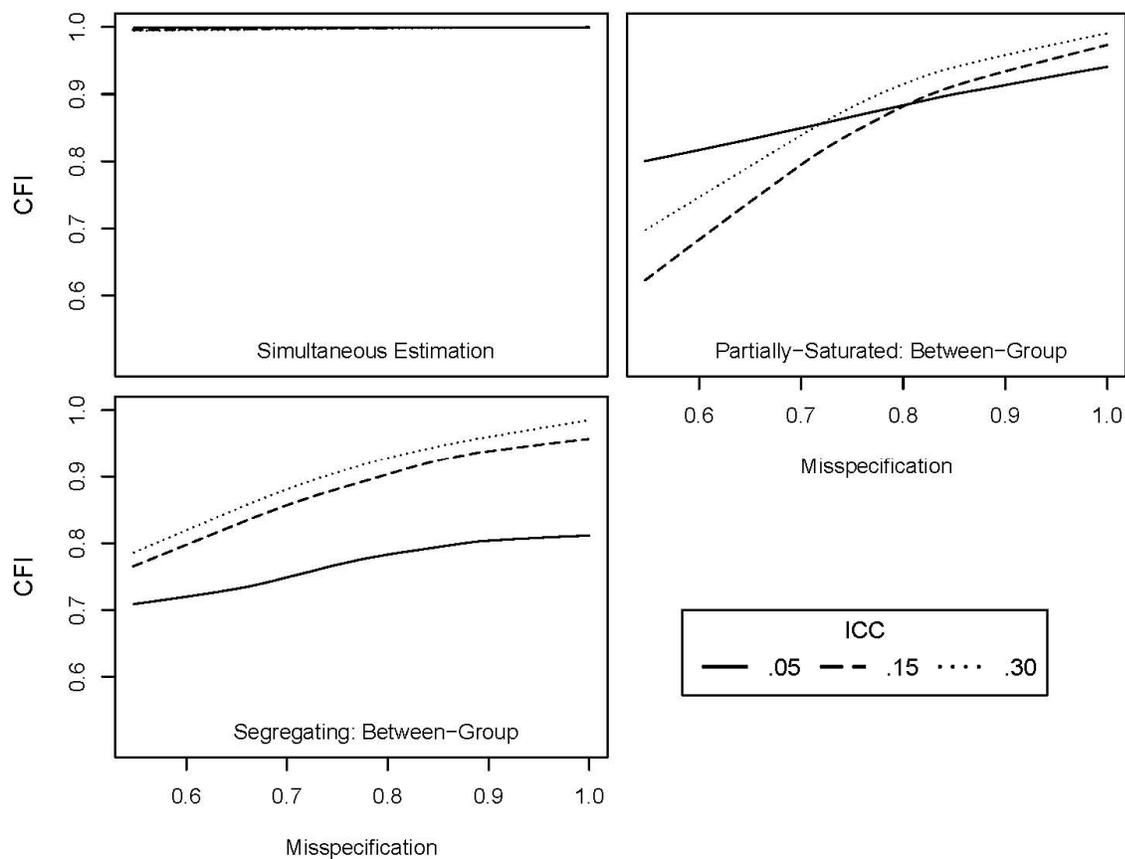


Figure 7. Loess Curves for CFI.

Note: Misspecification is defined on the x-axis by the value of the latent correlation  $\psi_{2,1}$ . A higher correlation value corresponds to less model misspecification.

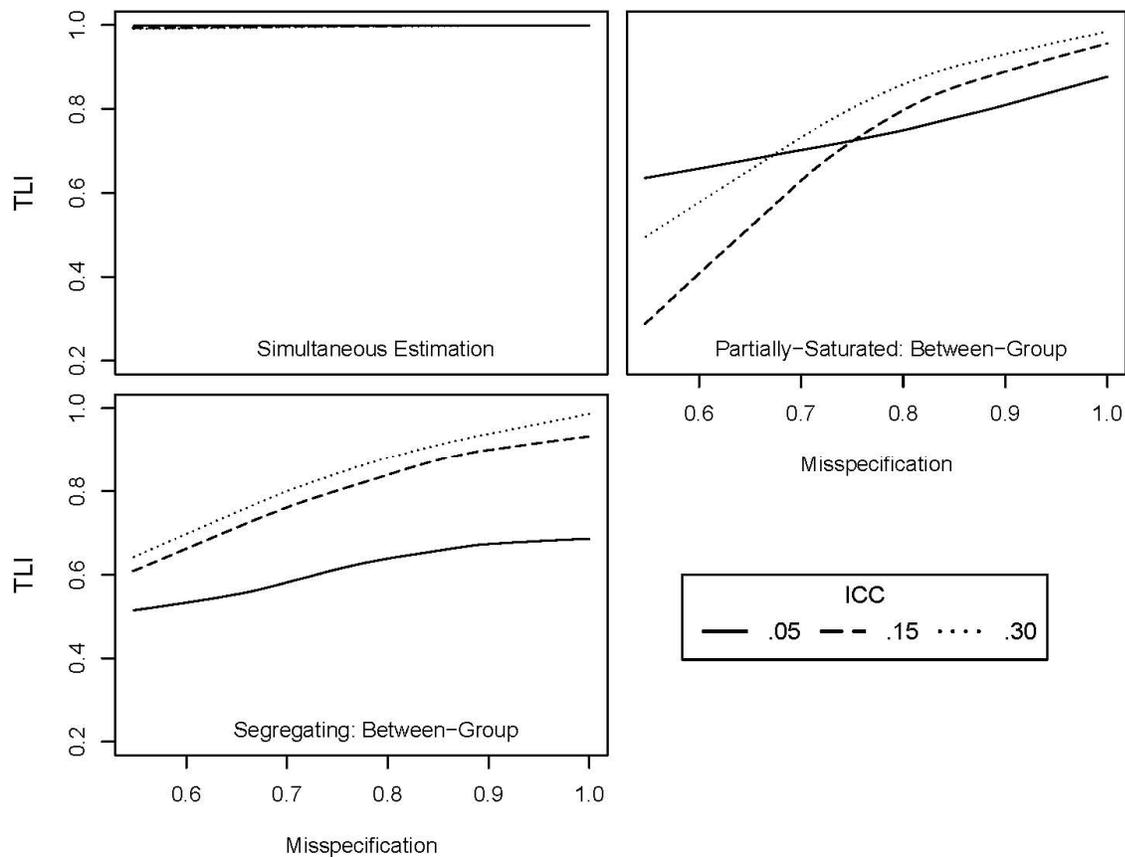


Figure 8. Loess Curves for TLI.

Note: Misspecification is defined on the x-axis by the value of the latent correlation  $\psi_{2,1}$ . A higher correlation value corresponds to less model misspecification.

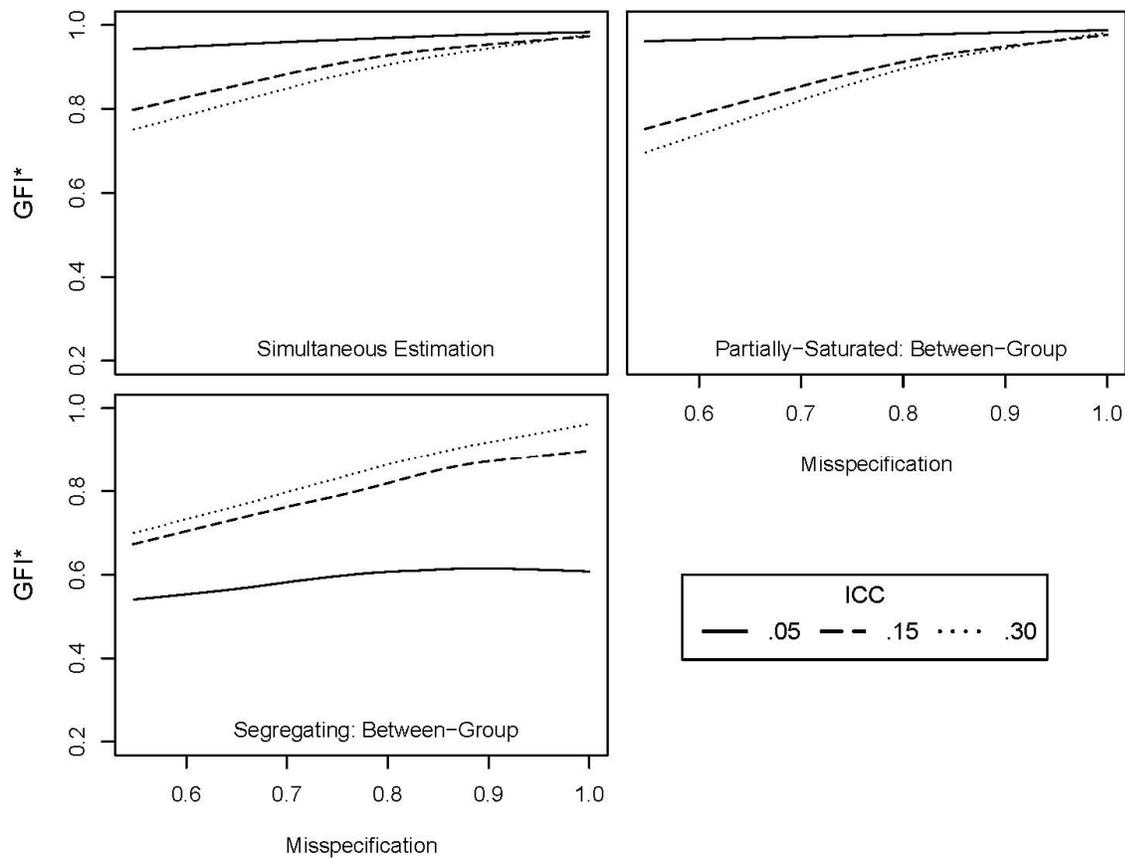


Figure 9. Loess Curves for GFI\*.

Note: Misspecification is defined on the x-axis by the value of the latent correlation  $\psi_{2,1}$ . A higher correlation value corresponds to less model misspecification.

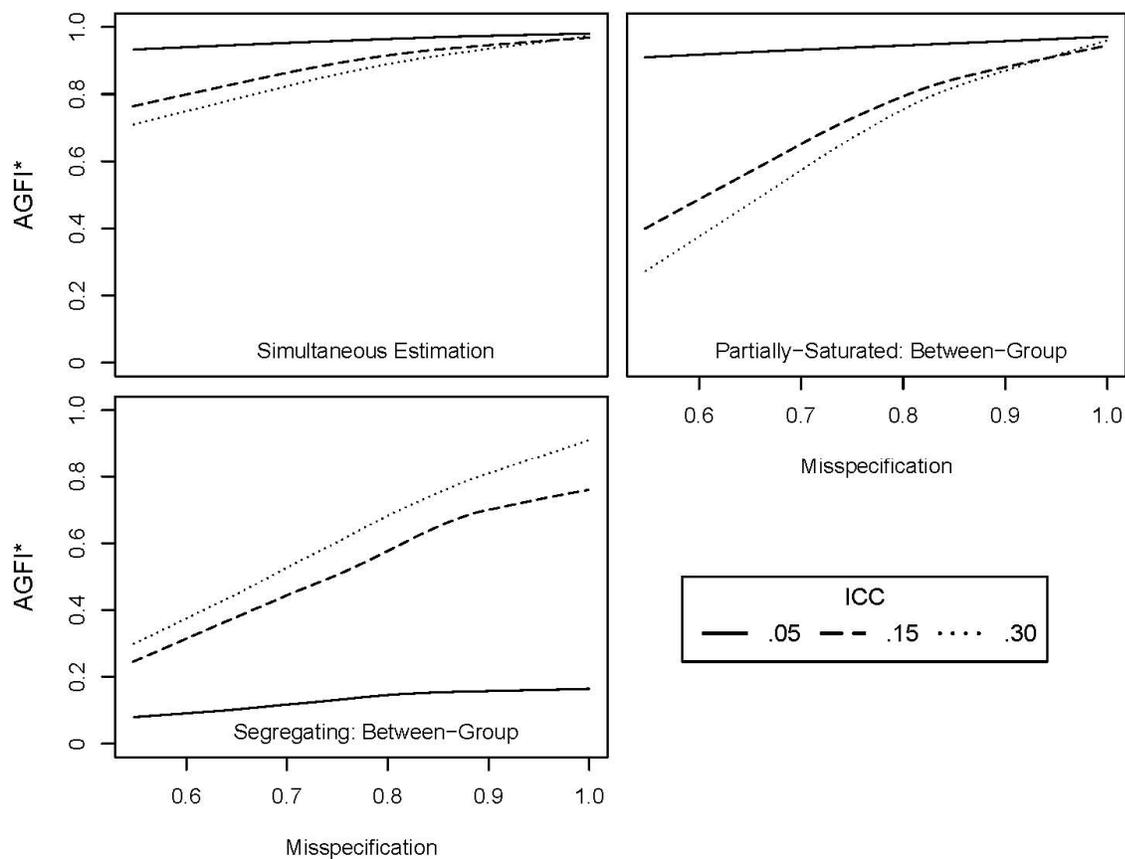


Figure 10. Loess Curves for AGFI\*.

Note: Misspecification is defined on the x-axis by the value of the latent correlation  $\psi_{2,1}$ . A higher correlation value corresponds to less model misspecification.

## APPENDIX A: R Program for Segregating Approach Implementation

```

#####
## Define seg.vech() function: Vectorizes the lower triangle of a matrix
#####
seg.vech <- function(A=0) {
  l <- 0;
  p <- nrow(A);
  pstar <- p*(p+1)/2;
  Va <- matrix(0,nrow=pstar,ncol=1,byrow=T);
  for (i in 1:p) {
    for (j in i:p) {
      l <- l+1
      Va[l] <- A[j,i]
    }
  }
  return(Va)
}
#####
## Define seg.DP() function: Creates a duplication matrix
#####
seg.DP <- function(p=0) {
  mat <- diag(p) ##Creates diagonal matrix of dimension p
  index <- seq(p*(p+1)/2)
  mat[ lower.tri( mat , TRUE ) ] <- index
  mat[ upper.tri( mat ) ] <- t( mat )[ upper.tri( mat ) ]
  outer(c(mat), index , function( x , y ) ifelse(x==y, 1, 0 ) )
}
#####
## Define seg.switch() function: Creates a permutation matrix
#####
seg.switch <- function(p=0) {
  ps <- p*(p+1)/2
  amat <- matrix(0,p,p)
  bmat <- matrix(0,p,p)
  na <- 0
  for (j in 1:p) {
    for (i in j:p) {
      na <- na+1
      amat[i,j] <- na;
    }
  }
  va <- seg.vech(amat)
  nb <- 0
  for (i in 1:p) {
    for (j in 1:i) {
      nb <- nb+1

```

```

        bmat[i,j] <- nb
      }
    }
  vb <- seg.vech(bmat)
  Imat <- diag(ps)
  permu <- Imat[va,vb]
  return(permu)
}
#####
## Define seg.data() function: Processes the data
#####
seg.data <- function(p=0,nlevel1=0,ng=0,yamat=0) {
  nt <- sum(nlevel1)
  ps <- p*(p+1)/2
  ymean <- matrix(0,nrow=ng,ncol=p)
  vsmatL1 <- matrix(0,nrow=ng,ncol=ps)
  smatw <- matrix(0,p,p)
  nsum <- 0
  for (jj in 1:ng) {
    nj <- nlevel1[jj,]
    ymatj <- yamat[(nsum+1):(nsum+nj),]
    nsum <- nsum+nj
    ymean[jj,] <- (matrix(1,1,nj)%*%ymatj)/nj
    njsmatj <- t(ymatj)%*%(diag(nj)-(matrix(1,nj,nj)/nj))%*%ymatj
    smatw <- smatw+njsmatj
    vsmatj <- seg.vech(njsmatj)
    vsmatL1[jj,] <- t(vsmatj)
  }
  smatw <- smatw/(nt-ng)
  output <- list(smatw,ymean,vsmatL1)
  return(output)
}
#####
## Define seg.mdm1() function: Evaluates mu0 and sig for saturated L2 model
#####
seg.mdm1 <- function(p=0,beta=0) {
  mu <- beta[1:p]
  sigw <- matrix(0,p,p)
  nc <- p
  for (i in 1:p) {
    for (j in i:p) {
      nc <- nc+1
      sigw[j,i] <- beta[nc]
      sigw[i,j] <- beta[nc]
    }
  }
}

```

```

vsigw <- seg.vech(sigw)
sigb <- matrix(0,p,p)
for (i in 1:p) {
  for (j in i:p) {
    nc <- nc+1
    sigb[j,i] <- beta[nc]
    sigb[i,j] <- beta[nc]
  }
}
vsigb <- seg.vech(sigb)
output <- list(mu,sigb,vsigb,sigw,vsigw)
return(output)
}
#####
## Define seg.minQ1() function: Maximizes the LL f() for L2 saturated model
#####
seg.minQ1 <- function(dup=0,nlevel1=0,ng=0,beta0=0,smatw=0,ymean=0,vsmatL1=0) {
  nt <- sum(nlevel1)
  p <- nrow(smatw)
  ps <- p*(p+1)/2
  ep <- .00001
  vsmatw <- seg.vech(smatw)
  err <- 0
  ## Gauss-Newton begins here
  iitera <- 0
  for (i in 1:51) {
    sigwb0 <- beta0[(p+1):(2*ps+p)]
    if (iitera>50) {
      err <- 1;
      write(cat("iterations=",iitera));
      stop("Maximum number of iterations exceeded")
    }
    iitera <- iitera+1
    output.mdm1 <- seg.mdm1(p,beta0)
    ## De-list
    mu <- output.mdm1[[1]]
    sigb <- output.mdm1[[2]]
    vsigb <- output.mdm1[[3]]
    sigw <- output.mdm1[[4]]
    vsigw <- output.mdm1[[5]]
    siginw <- solve(sigw)
  ## Weight given by normal theory
  weightw <- 0.5*t(dup)%*(siginw%x%siginw)%*%dup
  ssiginj <- matrix(0,p,p)
  ssymj <- matrix(0,p,1)
  gt2 <- matrix(0,ps,1)
}

```

```

gt3 <- matrix(0,ps,1)
ddljj <- matrix(0,ps,ps)
ddljjw <- matrix(0,ps,ps)
ddlww <- matrix(0,ps,ps)
for (j in 1:ng) {
  nj <- nlevel1[j]
  sigj <- sigb+sigw/nj
  vsigj <- seg.vech(sigj)
  siginj <- solve(sigj)
  weightj <- 0.5*t(dup)%*%(siginj%x%siginj)%*%dup
  ssiginj <- ssiginj+siginj
  ymj <- t(t(ymean[j,]))
  ssymj <- ssymj+siginj%*%ymj
  cymj <- ymj-mu
  Rj <- cymj%*%t(cymj)
  vrj <- seg.vech(Rj)
  cvrj <- vrj-vsigj
  wcvrj <- weightj%*%cvrj
  gt2 <- gt2+wcvrj/nj
  gt3 <- gt3+wcvrj
  ddljj <- ddljj+weightj
  ddljjw <- ddljjw+weightj/nj
  ddlww <- ddlww+weightj/(nj*nj)
}
mu1 <- solve(ssiginj)%*%ssymj
gt2 <- (nt-ng)*weightw%*%(vsmatw-vsigt)+gt2
gta <- rbind(gt2,gt3)
ddljjw <- t(ddljjw)
ddlww <- (nt-ng)*weightw+ddlww
ddl <- rbind(cbind(ddlww,ddljjw),cbind(ddljjw,ddljj))
stdi <- solve(ddl)
delt <- stdi%*%gta
sigwb1 <- sigwb0+delt
beta0 <- rbind(mu1,sigwb1)
dt <- sum(delt^2)/sum(sigwb1^2)
if (dt<ep) {
  output <- list(stdi,err);
  return(output)
}
}
return("Maximum number of iterations exceeded")
}
#####
## Define seg.Ascov() function: Maximizes the LL f() for L2 saturated model
#####
seg.Ascov <- function(beta0=0,p=0,dup=0,nlevel1=0,ng=0,smatw=0,ymean=0,vsmatL1=0) {

```

```

p <- nrow(smatw)
ps <- p*(p+1)/2
nt <- sum(nlevel1)
output.mdm1 <- seg.mdm1(p,beta0)
  ## De-list
  mu <- output.mdm1[[1]]
  sigb <- output.mdm1[[2]]
  vsigb <- output.mdm1[[3]]
  sigw <- output.mdm1[[4]]
  vsigw <- output.mdm1[[5]]
siginw <- solve(sigw)
weightw <- 0.5*t(dup)%*%(siginw%x%siginw)%*%dup
ddluu <- matrix(0,p,p)
ddljj <- matrix(0,ps,ps)
ddljjw <- matrix(0,ps,ps)
ddlww <- matrix(0,ps,ps)
Bmat <- matrix(0,(p+2*ps),(p+2*ps))
for (jj in 1:ng) {
  nj <- nlevel1[1]
  sigj <- sigb+sigw/nj
  siginj <- solve(sigj)
  vsigj <- seg.vech(sigj)
  ## Weight given by normal theory
  weightj <- 0.5*t(dup)%*%(siginj%x%siginj)%*%dup
  ddluu <- ddluu+siginj
  ddljj <- ddljj+ddljj+weightj
  ddljjw <- ddljjw+weightj/nj
  ddlww <- ddlww+weightj/(nj*nj)
  ymj <- t(ymean[1,])
  cymj <- ymj-mu
  gj1 <- siginj%*%cymj
  Rj <- cymj%*%t(cymj)
  vrj <- seg.vech(Rj)
  cvrj <- vrj-vsigj
  wcvrj <- weightj%*%cvrj
  gj3 <- wcvrj
  wcvswj <- weightw%*%(vsmatL1[1,]-(nj-1)*vsigw)
  gj2 <- wcvrj/nj+wcvswj
  gj <- rbind(gj1,gj2,gj3)
  Bmat <- Bmat+gj%*%t(gj)
}
ddljjw <- t(ddljjw)
ddlww <- (nt-ng)*weightw+ddlww
Amat <- rbind(cbind(ddluu,matrix(0,p,ps),matrix(0,p,ps)),
  cbind(matrix(0,ps,p),ddlww,ddljjw),
  cbind(matrix(0,ps,p),ddljjw,ddljj))

```

```

Amat <- Amat/ng
stdi <- solve(Amat)
Bmat <- Bmat/ng
Gamma <- stdi%*%Bmat%*%stdi
}
#####
## Main Program
#####
## Read in data
setwd("D:/Users/abouton/Desktop")
x <- read.table("rep1.dat") ##### DATA FILE NAME HERE
## Level-1 sample size
nt <- nrow(x)
## Level-2 sample size
xcl <- x[,7] ##### INSERT CLUSTER VARIABLE COLUMN HERE
ng <- 1
x1 <- xcl[1]
for (i in 2:nt) {
  xi <- xcl[i]
  if (xi==x1) {ng <- ng+0}
  else {ng <- ng+1; x1 <- xi}
}
## Generates the level-1 sample size variable 'nlev1'
x1 <- xcl[1]
nlev1 <- matrix(0,ng,1)
jj <- 1
for (i in 1:nt) {
  if (xcl[i]==x1) {nlev1[jj] <- nlev1[jj]+1}
  else {jj <- jj+1; nlev1[jj] <- nlev1[jj]+1; x1 <- xcl[i]}
}
## Data preparation
x2 <- as.matrix(x)
ymat <- as.matrix(x[,1:6])##### INSERT DEPENDENT VARIABLE COLUMNS HERE
p <- ncol(ymat)
write(cat("number of variables=",p,"\n"))
ps <- p*(p+1)/2
ybar <- matrix(1,1,nt)%*%ymat/nt
smat <- t(ymat)%*%(diag(nt)-matrix(1,nt,nt)/nt)%*%ymat/nt
vsmat <- seg.vech(smat)
beta00 <- rbind(t(ybar),(vsmat/2),(vsmat/2))
nlevel1 <- nlev1
nt <- sum(nlevel1)
nbar <- nt/ng
dup <- seg.DP(p)

## Prepares data and calculates the sample means and covariances

```

```

output.data <- seg.data(p,nlevel1,ng,yamat)
  ## De-list
  smatw <- output.data[[1]]
  ymean <- output.data[[2]]
  vsmatL1 <- output.data[[3]]
beta0 <- beta00
output.minQ1 <- seg.minQ1(dup,nlevel1,ng,beta0,smatw,ymean,vsmatL1)
  ## De-list
  stdi <- output.minQ1[[1]]
  err <- output.minQ1[[2]]
if (err==0) {output.Ascov <- seg.Ascov(beta0,p,dup,nlevel1,
  ng,smatw,ymean,vsmatL1)}
  ## De-list
  Amat <- output.Ascov[[1]]
  Gamma <- output.Ascov[[2]]
Gamma11 <- (nbar-1)*Gamma[(p+1):(ps+p),(p+1):(ps+p)]
Gamma22 <- Gamma[(ps+p+1):(2*ps+p),(ps+p+1):(2*ps+p)]
Gamma12 <- sqrt(nbar-1)*Gamma[(p+1):(ps+p),(ps+p+1):(2*ps+p)]
output.mdm1 <- seg.mdm1(p,beta0)
  ## De-list
  mu <- output.mdm1[[1]]
  sigb <- output.mdm1[[2]]
  vsigb <- output.mdm1[[3]]
  sigw <- output.mdm1[[4]]
  vsigw <- output.mdm1[[5]]
permu <- seg.switch(p)
print("-----")
print("-----")
Nw <- nt-ng
cat("the sample size equivalent number N-J for analyzing level-1 alone=",
  Nw, sep = "")
sbigw <- sigw
vsw <- vsigw
print("\hat\Sigma_1=")
print(sbigw)
print("-----")
print("-----")
Gamma11 <- permu%%Gamma11%%t(permu) # This line is not necessary if the
# weight matrix needs to be
# in the order of vech(\hat\Sigma);

print("\hat\Gamma11=")
print(Gamma11)
print("-----")
print("-----")
print("-----")
print("-----")

```

```
sbigb <- ng*sigb/(ng-1)
vsb <- vsigb
print("\hat\Sigma_2=")
print(sbigb)
print("-----")
print("-----")
Gamma22 <- permu%*%Gamma22%*%permu
print("\hat\Gamma22=")
print(Gamma22)
print("-----")
print("-----")
```