July 2019

Statistical Models to Test Measurement Invariance with Paired and Partially Nested Data: A Monte Carlo Study

Diep Thi Nguyen
University of South Florida, diepdrm@gmail.com

Statistical Models to Test Measurement Invariance with Paired and Partially Nested Data: A Monte Carlo Study

by

Diep Thi Nguyen

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Educational Measurement and Research
College of Education
University of South Florida

Co-Major Professor: Eunsook Kim, Ph.D.
Co-Major Professor: John M. Ferron, Ph.D.
Robert Dedrick, Ph.D.
Tony Tan, Ed.D.

Date of Approval:
June 14, 2019

Keywords: multilevel modeling, measurement equivalence, simulation, nested data

ACKNOWLEDGEMENTS

I am so fortunate to know many great people who without their support, guidance and encouragements, my Ph.D. journey could not be fulfilled with lots of joy and accomplishments. First I would like to express my sincere thanks to my former advisor and first mentor until he retired, Dr. Jeffrey Kromrey. He was the professor who first taught me statistics in education as well as how to do research and utilize SAS programming to statistical analyses and the person who always encouraged and created opportunities for his students to engage in doing research.

I am really grateful to have support and guidance from my wonderful dissertation committee. I would like to send my heartfelt thanks to my major advisor, Dr. Eunsook Kim who without her teaching and help I could not being a researcher as I am now. She is always my role model as a researcher and teacher. In the same manner, my special thanks also go to Dr. Ferron, my co-major advisor. I really appreciate and enjoy so much all the time he has spent with me to discuss about all issues of statistics, teaching and consulting and I cannot count how much I have learned and get inspired from him.

I would like to thank Dr. Robert Dedrick and Dr. Tony Tan so much for their continuing support and valuable advice not only for my dissertation but also other measurement and psychology studies. Their prompt responses even during the weekend were highly appreciated. I also would like to send my sincere thanks to Drs. Liliana Rodriguez-Campos, Shannon Suldo, Elizabeth Shaunessy-Dedrick, and Kathy Bradley-Klug who always cared, encouraged and infused positive energy to me. I was lucky to be a part of the Educational Measurement and Research family at the University of South Florida with a truly supportive, friendly and

collegiate environment. I am thankful to receive support from all faculty and staff of this program as well as to know and share exciting moments with fellow colleagues (Chunhua Cao, Eun Kyeng Baek, Yan Wang, Patricia Rodriguez deGil and others). I particularly thank Dr. Ha Phan who introduced me to this field and instilled her enthusiasm to me.

I am in debt with strong support, trust and unconditional love from my family and friends. No words can convey my gratitude and thanks to my parents who and sacrificed and worked extremely hard so I could get the best education, who raised me with full of love, care and beliefs, and from whom I inherited not only the ‘never give up, you can do it’ motto but also the feeling of happiness and love while helping other people. I am really grateful to have continuing support, trust, and love from my wonderful husband and two sons, especially when I was busy and had to come back home late. I am especially thankful to my sister and brother in law for always loving, believing, helping and being by my side. Last but not least, I would like to greatly thank my friend Ai Hoang, and my best childhood friends in Vietnam who despite a half of the earth distance from me but were always close to me, for their love, beliefs and encouragements.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

While assessing emotions, behaviors or performance of preschoolers and young children, scores from adults such as parent psychiatrist and teacher ratings are used rather scores from children themselves. Data from parent and teacher ratings are often nested such as students are within teachers and a child is within their parents. This popular nested feature of data in social, behavioral and health sciences makes measurement invariance (MI) testing across informants of children methodologically challenging. There was lack of studies that take into account the nested structure of data in MI testing for multiple adult informants, especially no simulation study that examines the performance of different models used to test MI across different raters.

This dissertation focused on two specific nesting data types in testing MI between adult raters of children: paired and partial nesting. For the paired data, the independence assumption of regular MI testing is often violated because the two informants (e.g., father and mother) rate the same child and their scores are anticipated to be related or dependent. Thus, in case of teacher and parent ratings of the same children, data are repeated measures and also partially nested. I proposed and evaluated the performance of the two statistical models that can handle repeated measures and partial nesting with several simulated research scenarios in addition to one commonly used and one potentially appropriate statistical model across several research scenarios. Results of the two simulation studies in this dissertation showed that for the paired data, both repeated measure confirmatory factor analysis (CFA) and multiple-group CFA models (Model 1 and Model 2, respectively) were able to detect scalar invariance most of the time using $\Delta\chi 2$ test and ΔCFI with higher rates for multiple-group CFA model. For configural invariance

and metric invariance conditions for the paired data, Model 1 had higher detection rates than Model 2 in almost research scenarios examined in Study 1. Particularly while Model 1 could detect noninvariance (either in intercepts only or in both intercepts and factor loadings) than Model 2 for paired data most of the time, Model 2 could rarely catch it if using suggested cut-off of 0.01 for RMSEA difference. For the paired data, Model 1 might be favored if researchers are more interested in detecting noninvariance due to its overall high detection rates for configural and metric levels of MI and Model 2 could be a good choice if the focus is on testing scalar invariance. For scalar invariance with partially nested data, both multilevel repeated measure CFA (Model 3) and design-based multilevel CFA (Model 5) could detect invariance in many conditions (from 81% to 100% of examined cases) with slightly higher detection rate for the former model than the later. Multiple-group CFA model (Model 4) could hardly detect scalar invariance except when ICC was small. The detection rates for configural invariance using $\Delta\chi 2$ test or Satorra-Bentler likelihood ratio test were also highest for Model 3 (82% to 100% except only two conditions with detection rates of 61%), following by Model 5 and the lowest was Model 4. Models 4 and 5 could reach these rates only with the largest sample sizes (i.e., large number of cluster or large cluster size or large in both factors) when the magnitude of noninvariance was small. Unlike scalar and configural invariance, the ability to detect metric invariance was highest for Model 4, following by Model 5 and lowest for Model 3 across many conditions using all of the three performance criteria. As higher detection rates for all configural and scalar invariance, and moderate detection rates for many metric invariance conditions (except cases of small number of clusters combined with large intraclass correlation, ICC), Model 3 could be a good candidate to test measurement invariance with partially nested data when having sufficient number of clusters or if having small number of clusters with small ICC.

Model 5 might be also a reasonable option for this type of data if both the number of clusters and cluster size were large (i.e., 80 and 20, respectively), or either one of these two factors was large coupled with small ICC. If ICC is not small, it is recommended to have a large number of clusters or combination of large number of clusters and large cluster size to ensure high detection rates of measurement invariance for partially nested data. As multiple group CFA had better and reasonable detection rates than the design-based and multilevel repeated measure CFA models with the conditions of small cluster size (10) coupled with small ICC (0.13) across configural, metric and scalar invariance levels, researchers can consider using this model to test measurement invariance when they can only collect 10 participants within a cluster (e.g. students within a classroom) and the degree of data dependency is small (e.g. small variance between clusters).

CHAPTER ONE: INTRODUCTION

1. Problem Statement

Children's emotions, behaviors, cognitions, or performance is often assessed using scores from multi-item scales or surveys on latent constructs. Scores from multiple informants (e.g., parents and teachers, mothers and fathers, patients and doctors) are widely used in research in education, psychology, and health sciences for many research purposes such as construct validation, theory development, or predicting outcomes. For example, McCarthy's dissertation (2015) used T-scores of parent and teacher behavioral and emotional assessments via ratings on the BASC-2 Behavioral and Emotional Screening System (BESS) instrument developed by Kamphaus and Raynolds (2007) to predict students' math achievement. In another study, Kenny, Veldhuijzen, Van Der Weijden, LeBlanc, Lockyer, Légaré, and Campbell (2010) compared the degree of agreement between doctors and patients as well as among patients of the same doctors on doctor communication skills using the same Matched-Pair Instrument. According to Kraemer, Measelle, Ablow, Essex, Boyce, and Kupfer (2003) and De Los Reyes and colleagues (2015), multiple informant assessment is one of the most widely used approaches to evaluate contextual variations in mental health. Ratings from informants based on their observations of patients' behavior in specific contexts (e.g., home vs. school vs. peer interactions) can provide information on how consistently or inconsistently concerns about patients' behavior are presented across those settings (Dirks, Reyes, Briggs-Gowan, Cella, & Wakschlag, 2012; De Los Reyes et al., 2015).

When using ratings of children or patients from multiple informants, it is of concern whether or not the scores from these reporters can be compared and interpreted in the same way within and across studies, which brings up the issues of measurement invariance (MI), also called measurement equivalence. For example, is the dimensionality of children's emotional and behavioral problems (e.g., internalizing concerns, externalizing concerns, personal problems) investigated with parent ratings in one study comparable to that investigated with teacher ratings in another study? Or do mother and father responses to items about child aggression mean the same thing? These questions about the invariance of a test between different informants of children are far-reaching because the test scores are often used for intervention choice or placement of children in a certain treatment program. Therefore before using scores of different resources such as parent and teacher ratings of a certain assessment, it is necessary to examine whether these parents and teachers perceive the underlying constructs in that instrument in the same manner. Similarly, in order to compare scores of doctors and patients or use these scores to predict an outcome, one needs to make sure these groups of informants interpret the items in the assessment equally, or MI of the instrument is achieved. As a result of MI, the same construct is measured across informants and comparisons between informants are meaningful. In other words, any observed differences between informants can be attributed solely to the measure of interest rather than artifact effects.

The importance of establishing MI when using scale scores among multiple groups is emphasized in the Standards (AERA, APA, & NCME, 2014). However, issues of MI among different informants using the same assessment have not received much attention in the literature. Among applied studies that performed MI across raters (e.g., Konold, Walthall, & Pianta, 2004; Woehr, Sheehan, & Bennett, 2005; Burns, Servera, del Mar Bernard, Carrillo, &

Geiser, 2013), most of them used multiple-group confirmatory factor analysis (CFA) to compare the measurement models between those informants. Studies by Burns, de Moura, Walsh, Desmul, Silpakit, and Sommers-Flanagan (2008), Burns, Walsh, Servera, Lorenzo-Seva, Cardo, and Rodríguez-Fornells (2013), Clark, Listro, Durbin, Donnellan, and Neppl (2016), as well as Piskernik, Supper, and Ahnert (2018), were the few that used a CFA approach with the inclusion of correlations between informant factors as well as errors between identical items from these two factors while examining MI across informants. The current study advocates testing MI between multiple informants of children or participants as a standard procedure before the use of scale scores from these reporters.

While assessing emotions, behaviors, or performance of preschoolers and young children, scores from adults such as parent, psychiatrist, and teacher ratings are used rather than scores from the children themselves. Data from parent ratings or from parents and teachers are often nested such as students are within teachers and a child is within their parents. This popular nested feature of data in educational, social and behavioral sciences makes MI testing across informants of children methodologically challenging. While the importance of MI has long been addressed and recommended for general assessment (e.g., Borsboom, 2006; Meredith & Teresi, 2006; Meade & Bauer, 2007; Yoon & Millsap, 2007; Fan & Sivo, 2009) as well as in multilevel setting (e.g., Kim, Kwok, &Yoon, 2012; Jak, Oort, & Dolan, 2013; Ryu, 2014), there have been a lack of studies that take into account the nested structure of data in MI testing for multiple adult informants. Especially, there has been no simulation study that has examined the performance of different models used to test MI across different raters.

The two simulation studies in this dissertation focus on two specific data types in testing MI between adult raters of children: paired and partial nesting. For the paired data, the

independence assumption of regular MI testing is often violated because the two informants (e.g., father and mother) rate the same child and their scores are anticipated to be related or dependent. I refer to this type of data dependency as repeated measures. The partial nesting data refers to the research situation where teacher and parent ratings are compared. In this scenario, it is common that each parent has only one child to rate while each teacher has multiple children in their classroom. In other words, children are nested within teachers but children are singletons for parents. Thus, in the case of teacher and parent ratings of the same children, data are repeated measures and also partially nested. Because of these unique features of data, MI testing between adult informants of children requires statistical models that take into account different types of data dependency. I used and evaluated the performance of the two statistical models that can handle repeated measures and partial nesting with several simulated research scenarios. When data are dependent due to repeated measures (father rating and mother rating of the same child), repeated measures CFA can be used for MI testing between informants; when data are partially nested in addition to repeated measures (parent rating and teacher rating of the same child), a model with specific model specification for partially nested data using multilevel repeated measures CFA can be employed. As multiple-group CFA is frequently used to examine MI across adult informants and the design-based multilevel CFA approach is potentially appropriate for the partial nesting characteristics of parent and teacher ratings, I was also interested in evaluating the performance of these two approaches with the paired and partially nested data described above.

2. Purposes of the Dissertation

This dissertation has two goals: (1) propose two statistical models to test measurement invariance between adult informants of children (e.g., father vs. mother, parent vs. teacher, etc.)

for ed and partially nested data, (2) examine the adequacy of the proposed models in comparison with the existing multiple-group CFA model used in the literature and the potential design-based multilevel CFA model through two Monte Carlo simulation studies.

There was a lack of studies that examine MI testing with paired and partially nested data in various research circumstances. In this dissertation, I conducted two simulation studies to investigate the two proposed MI models along with the commonly used multiple-group CFA model for paired and partially nested data as well as the design-based multilevel CFA model for partially nested data. To illustrate the theoretical framework for the two proposed models and other models examined in this dissertation, I present an example of the emotional construct (factor) of Inattention/Hyperactivity (in BESS, Kamphaus & Raynolds, 2007) with five items as the factor model that was used to conduct MI testing for each study.

3. Study 1: Repeated measures confirmatory factor analysis (CFA) for testing MI with paired data

The research setting for Study 1 was based on paired data. For this type of data, each informant (i.e., mother or father) assesses only one child or participant and both informants assess the same child or participant.

3.1. Model 1: Repeated Measures CFA (the Proposed Model)

For paired data, although children or participants are independent of each other, each of them has two sets of scores that are not independent. As both the mother and father give scores on the same child, the data can be considered as repeated measures although there is no time point (Olsen & Kenny, 2006). For this reason, I proposed the repeated measures CFA approach to testing the longitudinal MI as illustrated in Figure 1 for testing MI with paired data.

As presented in Figure 1, this CFA model consists of two factors with two identical sets of five items for each factor. In this model, because the scores from informants are often correlated to each other, the correlation or covariance between the two factors (i.e., mother and father factors of the child's emotional construct) is freely estimated. Importantly, the errors of each equivalent item from two informants are allowed to be correlated to reflect the data dependency between two scores from the two informants. The idea of incorporating correlation between mother and father factors as well as between unique factors of identical items was already mentioned in the CFA model for interchangeable dyads in Olsen and Kenny (2006, see their Figure 2 on page 131). However the purpose of their model was to perform SEM analysis rather than conduct MI testing for paired data to ensure the scores from the dyad are comparable or valid to use for other statistical analyses. In addition, their CFA model only applies to the measures where MI holds because in that model, factor loadings, intercepts, and error variances and item intraclass error covariances between mother and father factors were equal.

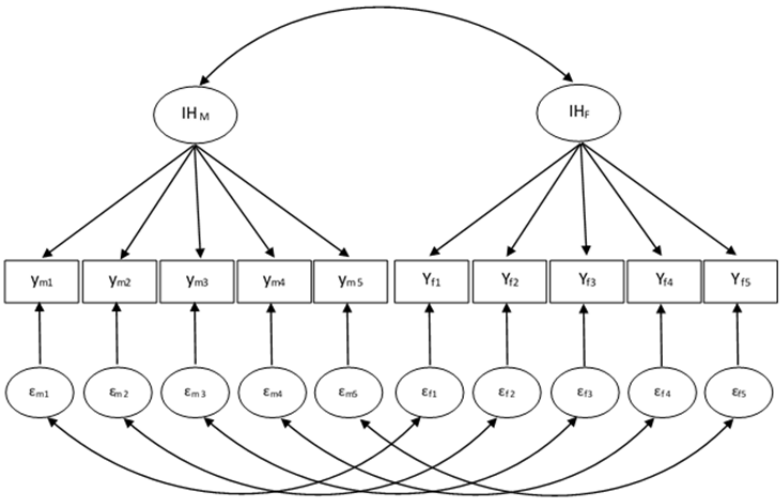


Figure 1. Repeated measures confirmatory analysis for mother (M) and father (F) factors of children's Inattention/Hyperactivity (IH) behaviors. For simplicity, the intercept of each item is not shown.

Similar to the general CFA model introduced in Kaplan (2009), the relationship between the two common factor scores (η_m for mother factor and η_f for father factor) and the continuous observed scores (y_{m1} to y_{m5} and y_{f1} to y_{f5}) in this structural equation model can be specified in a matrix format as follows:

$$\mathbf{Y} = \boldsymbol{\tau} + \boldsymbol{\Lambda\eta} + \boldsymbol{\varepsilon} \qquad (1)$$

where $\mathbf{Y}$ is a vector of observed variables, $\boldsymbol{\tau}$ is a vector of intercepts, $\boldsymbol{\Lambda}$ is matrix of factor loadings, $\boldsymbol{\eta}$ is a vector of two common factors (η_m and η_f), and $\boldsymbol{\varepsilon}$ is a vector of unique variables (ε_{m1}- ε_{m5} and ε_{f1}- ε_{f5}).

Equation (1.1) can be written as matrix forms with details as below:

$$\begin{pmatrix} y_{m1} \\ y_{m2} \\ y_{m3} \\ y_{m4} \\ y_{m5} \\ y_{f1} \\ y_{f2} \\ y_{f3} \\ y_{f4} \\ y_{f5} \end{pmatrix} = \begin{pmatrix} \tau_{m1} \\ \tau_{m2} \\ \tau_{m3} \\ \tau_{m4} \\ \tau_{m5} \\ \tau_{f1} \\ \tau_{f2} \\ \tau_{f3} \\ \tau_{f4} \\ \tau_{f5} \end{pmatrix} + \begin{pmatrix} \lambda_{m1} & 0 \\ \lambda_{m2} & 0 \\ \lambda_{m3} & 0 \\ \lambda_{m4} & 0 \\ \lambda_{m5} & 0 \\ 0 & \lambda_{f1} \\ 0 & \lambda_{f2} \\ 0 & \lambda_{f3} \\ 0 & \lambda_{f4} \\ 0 & \lambda_{f5} \end{pmatrix} \text{x} \begin{pmatrix} \eta_m \\ \eta_f \end{pmatrix} + \begin{pmatrix} \varepsilon_{m1} \\ \varepsilon_{m2} \\ \varepsilon_{m3} \\ \varepsilon_{m4} \\ \varepsilon_{m5} \\ \varepsilon_{f1} \\ \varepsilon_{f2} \\ \varepsilon_{f3} \\ \varepsilon_{f4} \\ \varepsilon_{f5} \end{pmatrix}$$

Based on the assumption that common factors and unique factors are uncorrelated, the covariance structure can be specified as:

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\,\boldsymbol{\Phi}\,\boldsymbol{\Lambda}' + \boldsymbol{\Theta}_{\boldsymbol{\varepsilon}} \qquad (2)$$

where $\boldsymbol{\Sigma}$ is a population covariance matrix, $\boldsymbol{\Phi}$ is a variance covariance matrix for common factors (latent variables), and $\boldsymbol{\Theta_\varepsilon}$ is a variance covariance matrix for the unique factors. Because this model (Model 1) allows the unique factors for identical items between mother and father factors to be correlated, the variance covariance matrix of the unique factors ($\boldsymbol{\Theta_\varepsilon}$) is different from that of regular CFA models where the unique factor correlations are all assumed to be zero. The common factor variance covariance matrix is shown below:

$$\boldsymbol{\Phi} = \begin{bmatrix} \Phi_m & \\ \Phi_{fm} & \Phi_f \end{bmatrix}$$

It should be noted that in the common factor variance covariance matrix for Model 1, the variances of mother factor (Φ_m) and father factor 2 (Φ_f) are similar to those factor variances in multiple-group CFA models. However, the factor covariance between father and mother factors (Φ_{fm}) does not exist in multiple-group CFA approach for testing MI because the mother and father factors are estimated separately in mother and father data sets and there is no correlation between these two factors.

In the variance covariance matrix for unique factors below, the non-diagonal elements (i.e., covariances of two unique factors) are all zero in multiple-group CFA models. However, as the unique factors of identical items are allowed to be correlated in Model 1, these covariances of unique factors are expected to be non-zero as shown in the $\boldsymbol{\Theta_\varepsilon}$ matrix above. For simplicity, only the covariances of paired, identical items between mother and father ratings are shown and the upper part above the diagonal is identical with the lower part in this $\boldsymbol{\Theta_\varepsilon}$ matrix.

$$\boldsymbol{\Theta_{\varepsilon}} = \begin{bmatrix} \text{var}(\varepsilon_{m1}) & & & & & & & & & \\ & \text{var}(\varepsilon_{m2}) & & & & & & & & \\ & & \text{var}(\varepsilon_{m3}) & & & & & & & \\ & & & \text{var}(\varepsilon_{m4}) & & & & & & \\ & & & & \text{var}(\varepsilon_{m5}) & & & & & \\ \text{cov}(\varepsilon_{m1,}\ \varepsilon_{f1)} & & & & & \text{var}(\varepsilon_{f1}) & & & & \\ & \text{cov}(\varepsilon_{m2,}\ \varepsilon_{f2)} & & & & & \text{var}(\varepsilon_{f2}) & & & \\ & & \text{cov}(\varepsilon_{m3,}\ \varepsilon_{f3)} & & & & & \text{var}(\varepsilon_{f3}) & & \\ & & & \text{cov}(\varepsilon m_{4,}\ \varepsilon_{f4)} & & & & & \text{var}(\varepsilon_{f4}) & \\ & & & & \text{cov}(\varepsilon_{5,}\ \varepsilon_{10)} & & & & & \text{var}(\varepsilon_{f5}) \end{bmatrix}$$

Based on another assumption that the expected value of a unique variable is zero, $E(\varepsilon) = 0$, and the fact that expected value of a common factor is denoted as $E(\eta) = \kappa$, the mean structure of a general CFA model from Equation 1.1 can be as below:

$$\boldsymbol{E(Y) = \tau + \Lambda K} \quad (3)$$

where $\mathbf{E}(Y)$ is a vector for expected values of observed variables, $\boldsymbol{\tau}$ is a vector of intercepts, $\boldsymbol{\Lambda}$ is the matrix for factor loadings, and $\mathbf{K}$ is a vector of factor means.

For MI testing, the following three steps were applied to run statistical analyses with the whole combined dataset of mother and father responses:

(1) configural invariance: a CFA model with two factors is constructed: one factor based on father responses and the second factor based on mother responses. The factor loadings and intercepts of all items (except the reference item where loading is fixed to 1 for identification purpose) in the two factors are freely estimated;

(2) metric invariance can be tested by holding factor loadings of items between father and mother factors equally;

(3) scalar invariance can be further tested by holding intercepts between fathers and mothers, in addition to factor loadings, for all items equally.

3.2. Model 2: Multiple Group CFA

As stated by Geiser, Burns, and Servera (2014), measurement structures across raters using the same instrument have not gained much attention in the literature. Among studies that tested MI between mother and father raters (e.g., Konold et al., 2004; Mayfield et al., 2018), multiple-group CFA approach was often used to perform MI testing between these informants. Thus I also include a model that employed a multiple-group CFA approach to test MI in addition to the two proposed models for paired data in Study 1 and partially nested data in Study 2.

For the multiple-group CFA approach, the one factor model of Inattention/Hyperactivity with five items (BESS, Kamphaus & Raynolds, 2007) is specified for mother group and father group, separately (see Figure 2). For simplicity, the intercept and error of each item are not shown in this figure. The one factor CFA model for mother and father is tested separately so mother and father factors as well as the similar item pair (e.g. item 1, y_m1, in the mother model

and item 1, y_f1, in the father model) are not correlated to each other, which is different from Model 1 in this study.

Derived from the general structural equation models defined in Equations 1 through 3, the multiple group CFA model can be specified by incorporating a group indicator:

$$\mathbf{Y}_g = \boldsymbol{\tau}_g + \boldsymbol{\lambda}_g\boldsymbol{\eta}_g + \boldsymbol{\varepsilon}_g \quad (4)$$

$$\boldsymbol{\Sigma}_g = \boldsymbol{\lambda}_g\boldsymbol{\Phi}_g\boldsymbol{\lambda}'_g + \boldsymbol{\Theta}_{g\varepsilon} \quad (5)$$

$$E(\mathbf{Y}_g) = \boldsymbol{\tau}_g + \boldsymbol{\lambda}_g\mathbf{K}_g \quad (6)$$

where subscript *g* is a group indicator (g=m for mother and g=p for father in Study 1 or $g = 1, 2, \ldots, G$ for general multiple-group CFA testing) and others are as defined in the previous section. When observed scores are assumed normally distributed, "factorial invariance holds if the conditional mean and variance covariance of observed scores given factor scores are independent of group membership (*g*)" (Kim, 2011).

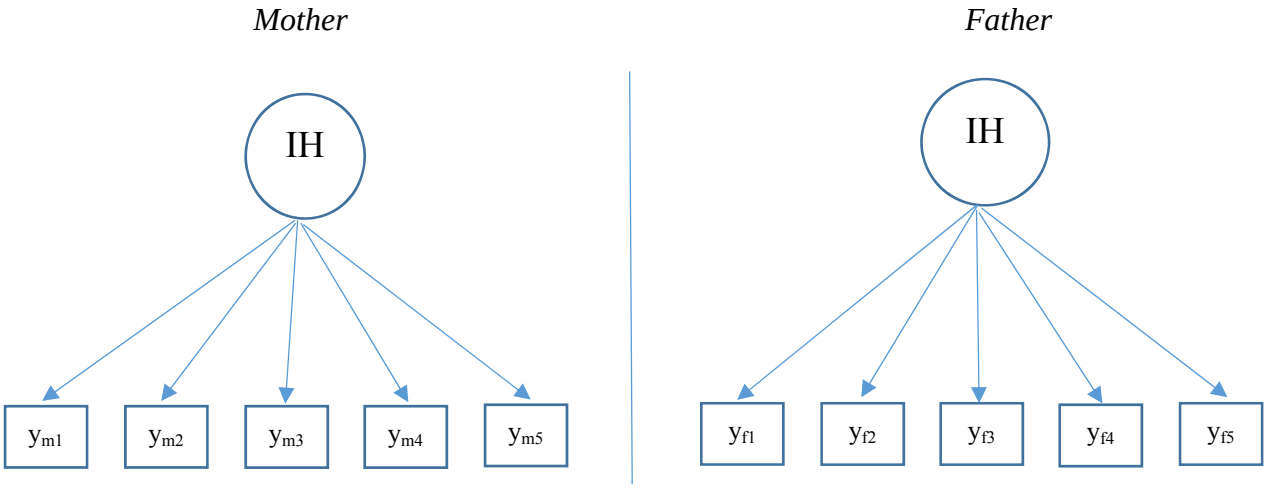


Figure 2. Multiple-group confirmatory analysis for mother (M) and father (F) factors of children's Inattention/Hyperactivity (IH) behaviors (BESS, Kamphaus & Raynolds, 2007)

The MI testing procedure for multiple-group CFA for this model also consists of three steps as in Model 1, i.e., from configural invariance, metric (or weak) invariance, and scalar (or strong) invariance. However the statistical model is the one factor with five items that is

constructed separately for mother and father as shown in Figure 2, which is different from the CFA model with two correlated factors as illustrated in Figure 1. Also, the data in this model (Model 2) are stacked up of mother and father responses and therefore the total sample size is doubled the number of children. The MI testing procedure for the one-factor CFA model is run separately for mother responses and father responses of children's assessments.

3.3. Research Questions for Study 1

To accomplish the study goals, I propose the following research questions:

1. How well does each model for the paired data detect the level of measurement invariance (configural, metric, or scalar invariance) under different research conditions?
2. What simulation design factors (e.g., factor correlations, degree of data dependency) are related to the performance of the proposed model as well as the comparative model for the paired data?

4. Study 2: Multilevel Repeated Measures CFA for Partially Nested Data

The research setting for Study 2 is based on partially nested data where one group of informants assesses only one child while the other group of informants assesses multiple children. Specifically, parents evaluate only their own child while the teachers evaluate multiple students in their classrooms. I propose a particular model (Model 3) for this type of data and also include the commonly used multiple-group CFA model (Model 4) and potentially appropriate design-based multilevel CFA model (Model 5) in Study 2.

4.1. Model 3: Multilevel Repeated Measures CFA (the Proposed Model)

For the partially nested data, alongside the repeated measure feature between parent and teacher ratings of the child, the data dependency also occurs in partial nesting where the child is nested within the teacher. The approaches for partially nested data using a structural equation

modeling framework suggested by Sterba et al. (2014) (e.g., a treatment group is multilevel, but a control group is single-level) may not be an optimal choice for this research scenario considering it does not take into account the additional data dependency (i.e., repeated measures). The commonly used multiple-group CFA in applied studies for testing MI with this type of partial nesting data will also not be appropriate because it does not consider the nested as well as repeated characteristics of the data.

In order to examine MI for this partially nested data, e.g., whether parents and teachers rate children in the similar manner, I propose the multilevel repeated measure CFA model. The statistical model is illustrated in Figure 3.

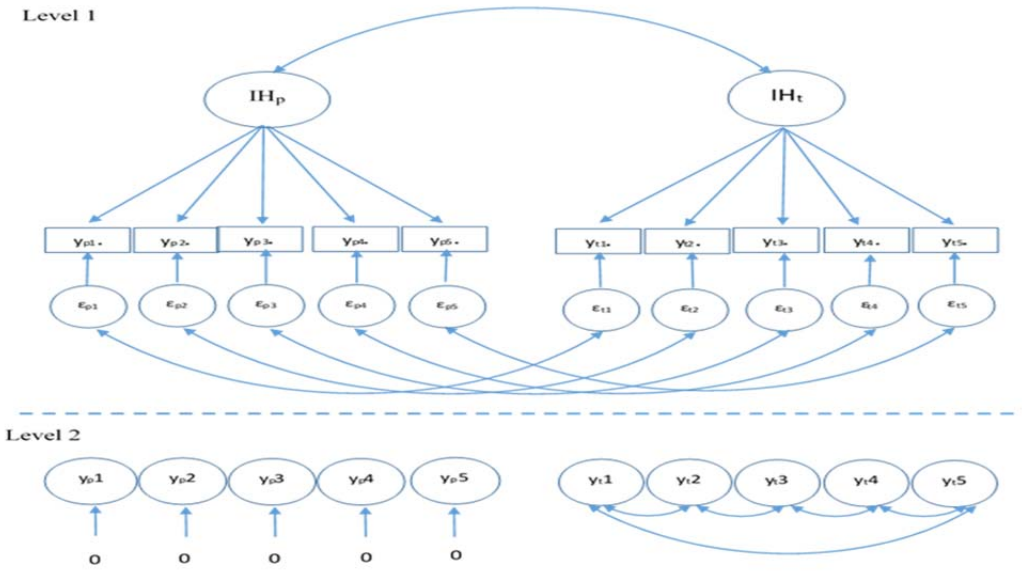


Figure 3. Multilevel repeated measures confirmatory factor analysis with partial nesting for Parent (IHp) and Teacher (IHt) factors of children's Inattention/Hyperactivity (IH) behaviors.

In this figure, the filled dots indicate random intercepts. At the between level, the variance and covariance of items y_{p1} through y_{p5} are all zeros while y_t1 through y_{t5} are all correlated to each other (only some drawn in the figure for simplicity). The intercept for each variable that is estimated at level 2 is also not shown for simplicity.

In this model, for within level, the unique factors and common factors between teachers and parents are correlated to reflect the data dependency from repeated measures, which is similar to Model 1. The variance covariance matrix for common factors ($\boldsymbol{\Phi}$) and unique factors ($\boldsymbol{\Theta}_{\varepsilon}$) are presented respectively as follows:

$$\boldsymbol{\Phi} = \begin{bmatrix} \Phi_p & \\ \Phi_{tp} & \Phi_t \end{bmatrix}$$

$$\boldsymbol{\Theta}_{\varepsilon} = \begin{bmatrix}
\text{var}(\varepsilon_{p1}) & & & & & & & & & \\
 & \text{var}(\varepsilon_{p2}) & & & & & & & & \\
 & & \text{var}(\varepsilon_{p3}) & & & & & & & \\
 & & & \text{var}(\varepsilon_{p4}) & & & & & & \\
 & & & & \text{var}(\varepsilon_{p5}) & & & & & \\
\text{cov}(\varepsilon_{p1,}\varepsilon_{t1}) & & & & & \text{var}(\varepsilon_{t1}) & & & & \\
 & \text{cov}(\varepsilon_{p2,}\varepsilon_{t2}) & & & & & \text{var}(\varepsilon_{t2}) & & & \\
 & & \text{cov}(\varepsilon_{p3,}\varepsilon_{t3}) & & & & & \text{var}(\varepsilon_{t3}) & & \\
 & & & \text{cov}(\varepsilon_{p4,}\varepsilon_{t4}) & & & & & \text{var}(\varepsilon_{t4}) & \\
 & & & & \text{cov}(\varepsilon_{p5,}\varepsilon_{t5}) & & & & & \text{var}(\varepsilon_{t5})
\end{bmatrix}$$

As shown in these matrices, unlike the multiple-group CFA approach where the non-diagonal elements in the variance covariance matrices of the unique factors and common factors are all zero or not available, the covariances of unique factors for each pair of identical items (e.g. $\text{cov}(\varepsilon_{p1}, \varepsilon_{t1})$) as well as the covariance between the two common factors (Φ_{tp}) are estimated in Model 3.

For the data dependency due to students nested within teachers, a multilevel modeling approach is adopted where students are at level 1 and teachers are at level 2. In order to take into account the partial nesting feature where parents are not nested within teachers, the variance covariance of parent scores are all constrained to be zero at the between level (level 2) while the variance-covariances of teacher scores are freely estimated at level 2. The constraint of zero for variance-covariances (analogous to intraclass correlation or ICC= 0) for parent scores illustrates no nesting feature for these variables.

For the within level in Study 2, the relationship between the observed scores of individuals (e.g., students) from ratings of their parents and teachers can be expressed by equations (7) and (8):

For the parent factor: $y_{pi} = \tau_p + \lambda_p \eta_{pi} + \varepsilon_{pi}$ (7)

where y_{pi} is the observed score for each item for student i from his/her parent, τ_p is the parent item intercept, λ_p is the factor loading of the parent item, η_{pi} is the parent common factor score for student i, and ε_{pi} is the unique factor of the parent item for student i.

For the teacher factor: $y_{tij} = \tau_{tj} + \lambda_{tj} \eta_{tij} + \varepsilon_{tij}$ (8)

where y_{tij} is the observed score for each item for student i from his/her teacher at a class j, τ_{tj} is the intercept and λ_{tj} is the factor loading of the teacher item at a class j, η_{tij} is the teacher common factor score for a student i at a class j, and ε_{tij} is the unique factor of the teacher item for a student i at a class j. It is assumed that factor loadings are invariant across classrooms in this study: $\lambda_{tj} = \lambda_t$.

It should be noted that τ_p and τ_{tj} are set equal to 0 in the within level because an individual score is its deviation from the group mean (Heck & Thomas, 2009; Kim & Cao, 2015).

In the between level in Study 2, there are only equations for intercepts for the parent and teacher factors because parents and teachers do not rate themselves and there are no factor structures at the between level. Because each parent rates only one child while one teacher often rates multiple children in his/her classroom, there are only variances in the teacher intercepts (τ_{tj}) but not in the parent intercepts (τ_p) at level 2 as shown in equations (9) and (10):

$$\tau_p = \mu_p \qquad (9)$$

where τ_p is the intercept of parent item and μ_p is the grand mean of intercepts from parent ratings for that item;

$$\tau_{tj} = \mu_t + \varepsilon_j \qquad (10)$$

where τ_{tj} is the intercept of the teacher item for class j, μ_t is the grand mean of the teacher item, ε_j is residual variance of teacher item for class j.

The MI procedure for this model is similar to the proposed Model 1 in Study 1. Specifically, first, configural invariance is tested using a CFA model with two factors of five items: one factor based on father responses and the second factor based on mother responses with no equality constraints of factor loadings and intercepts between the two factors. Then metric invariance can be tested by imposing equalities of factor loadings between parents and teachers. Third, scalar invariance testing is conducted by holding both factor loadings and intercepts equally between parents and teachers' scores.

4.2. Model 4: Multiple-group CFA

In their studies, Konold et al. (2004) as well as Waschbusch and Willoughby (2008) applied multiple-group CFA framework to examine the measurement equivalence between different informants in different settings (i.e., parents vs. teachers). Thus a single level multiple-group CFA approach is included in Study 2 as a comparative model with the proposed model.

This approach does not take into account the data dependency of parent and teacher ratings as well as the nested feature of students within teachers. The theoretical framework as well as MI testing sequence for this model are identical for those in Model 2 in Study 1 (see Figure 2).

4.3. Model 5: Design-based Multilevel CFA

For Study 2, because factors are not modeled at the between level, the design-based multilevel CFA approach (i.e., using the command TYPE = COMPLEX in Mplus program) can be an option used to address the data dependency although this approach might not be optimal for the partially nested data. The design-based multilevel CFA utilizes a single-level approach but takes into account the nested feature of data by adjusting the standard errors of the parameter estimates and the overall chi-square value (Muthén & Muthén, 1998–2010; Kim et al., 2012). Specifically, the one single-level model is specified with adjusted standard errors for sampling complexity instead of decomposing the covariance matrix into within and between components as in the regular multilevel CFA. Of note is that the adjusted standard errors are applied not only to teachers but also parents who have a single participating child. The common factors (parent and teacher factors) and the unique factors of identical items between the two common factors are correlated as in the within level (or level 1) of Model 3 (see Figure 3). As a result, variance covariance matrices of the common factors as well as unique factors of Model 5 are identical with those matrices for the within level of Model 3. The sequence of MI testing (i.e., from configural to scalar invariance) is the same as those hierarchical steps in Model 1 and Model 3.

4.4. Research Questions for Study 2

In order to accomplish the study goals, I examined the following research questions:

1. How well does each model detect the level of measurement invariance (configural, metric, or scalar invariance) under different research conditions for partially nested data?

2. What simulation design factors (e.g., sample size, degree of data dependency) are related to the performance of the proposed model as well as the comparative models for partially nested data?

5. Significance of the Research

Research about affect, cognition, and behaviors of children, especially preschoolers and young children, rely on data collected from different adult informants of children. As reported by Konold and Pianta (2007), "informant-based methods of behavioral assessment" are widely used and 90% of referrals for behavior problems of children in schools and medical settings are from behavior checklists in psychological assessments. However, studies about score alignment or agreement among multiple informants often focuses on evaluating how close or correlated the observed scores of these informants are. Little attention is paid to the question of whether or not different informants interpret and respond to survey items in the same manner and subsequently whether their scores can be comparable within and across studies. By investigating the score agreement from psychometric perspectives, this study emphasizes the importance of establishing MI before comparing observed scores of raters. This study also offers a means for educational and psychological researchers to investigate possible differences between raters in their perceptions, norms, and interpretations of children's affect, cognition, and behaviors as well as their response tendency or response styles.

Paired and partially nested data are common in educational, psychological and health sciences. Although multilevel modeling (Raudenbush & Bryk, 2002) that handles nested data and cross-classified data has been discussed and used extensively with the availability of numerous software programs, there has been a lack of studies about testing MI for paired and partially nested data. The two statistical models proposed and evaluated in this dissertation will

add to the MI literature a possible way to model data dependency with these special types of data. Results from the simulation studies provided practical guidelines for researchers in doing MI for their research. Moreover, the current study is expected to initiate further discussions on MI testing with other types of more complex data structures that possibly occur in educational, psychological and health sciences research.

CHAPTER TWO: LITERATURE REVIEW

This chapter reviews literature on studies about measurement invariance testing of psychological assessments using multiple informants. First, I discuss the issues on rater agreement of scores from multiple rating sources. Second I review applied studies conducting MI testing with multiple informants including inter-parental ratings and cross adult ratings. The methodological issues such as definition, levels and hierarchical procedures pertaining to measurement invariance testing will be introduced last.

1. Rater Agreement

Rater agreement is typically measured by comparing the means, evaluating the differences, or estimating the correlation between observed scores from multiple informants. Issues of such agreement or alignment among raters on children's attributes have been studied for decades. While information about the latent means is also often of interest, studies about multiple informant rating agreements usually focus on correlations between these ratings (Geiser et al., 2014). This correlation indicates "the degree of association between two variables scored for sets of variables (e.g. ASEBA, Achenbach System of Empirically Based Assessment, problem item scores) obtained from two individuals" (Achenbach & Rescorla, 2001, p. 34). Meta-analytic studies about adult rater agreements on children's problems have been conducted to examine the correlations between multisource ratings of psychological assessments. Achenbach, McConaughy, and Howell (1987) reviewed Pearson correlations between ratings of parents, teachers, and other informants from 269 samples in 119 studies. The agreement between mother and father ratings ($r = 0.6$) is reported to be higher than the agreement between parents

and teachers (r = .28) with higher discrepancies in the latter. These discrepancies between raters could possibly be due to parents and teachers observe children in different settings (e.g., Harvey, Fischer, Weieneth, Hurwitz, & Sayer, 2013). In subsequent meta-analytic studies, Duhig, Renk, Epstein, and Phares (2000) as well as De Los Reyes et al. (2015) also found that agreement between parental ratings was low to moderate. Specifically, Duhig et al. (2000) reported the average weighted correlations for maternal and paternal ratings of .45 for internalizing problems, .63 for externalizing problems and .70 for total behavior problems. De Los Reyes and colleagues (2015) reviewed cross-informant correspondences from 341 studies published between 1989 and 2014 that provided estimates of adult cross-informant correlations. Similar to results from Achenbach et al. (1987), the average correlations of cross-informants were found 0.25 for internalizing, 0.30 for externalizing and 0.28 for overall scale.

In addition to reporting correlations, studies about rater alignment also compare the latent means or assess the differences between observed scores from multiple informants. For example, Mayfield et al. (2018) evaluated the ratings equivalence of mother and father through examining the factor structure, means and correlations of the Diagnostic and Statistical Manual of Mental Disorders – Fourth Edition (DSM-IV) Attention Deficit and Hyperactivity Disorder (ADHD) Symptom Rating Scale. This instrument is used to assess different cognitive clusters in children diagnosed with ADHD symptoms. In their study, Mayfield and colleagues used intra-class correlation coefficient for absolute agreement to examine agreement between mother and father ratings for each of the three scales. They examined CFA models of mother and father, separately in terms of model fit (e.g. chi-square test, CFI, RMSEA, AIC). These authors also compared the means of constructs in the instrument between mother and father ratings to evaluate the equivalence of these ratings. However the MI across these informants was not conducted before

comparing the mean differences between mother and father scores. While the CFA models of mother and father ratings for this scale in Mayfield et al. (2018)'s study showed good fit for the three-factor model, it does not guarantee the MI between mother and father ratings satisfied.

Allegedly, literature about using multisource ratings of children's psychological assessments provides evidences for substantial misalignment of observed scores from different informants. However, there were few studies that examined the comparability of scores from multiple adult raters under psychometric perspectives in which the underlying construct of a measure rather than observed scores and differences in raters' perceptions, interpretations, and response styles of items are examined.

2. Review of Applied Studies that Examined Measurement Invariance across Informants

Although examining MI is considered as novel aspect for meaningful comparison across raters as well as providing significant information on the interpretation of mean rater effects (Geiser et al., 2014), there were only some studies that formally performed MI testing across multiple adult informant ratings.

Konold et al. (2004) examined possible explanation for discrepancies between multiple informants' ratings of the same child using the Child Behavior Checklist for Ages 4-18 (CBCL/4-18) and the Teacher Response Form (TRF) developed by Achenbach and Rescorla (2001). The CBCL/4-18 comprises 118 behavior problem items of which 85 items combine to form the eight narrow-band scales: Withdrawn, Somatic Complaints, Anxious/Depressed, Social Problems, Thought Problems, Attention Problems, Delinquent Behavior, and Aggressive Behavior and 93 items have counterparts on the TRF. Furthermore, five out of these eight narrow-band scales formed the two frequently used overarching broad-band scales: Internalizing

with combination of Withdrawn, Somatic Complaints, and Anxious/Depressed and Externalizing with Delinquent Behavior and Aggressive Behavior. The authors used multiple-group CFA to compare the two measurement models across informants (i.e. mothers vs. fathers with N=710 children, and parents vs. teachers with N=562): a configural model with no constraints of parameters (loadings, intercepts, variances) and a model that constrained equal factor loadings, equal error variances and equal factor correlations. The comparison of intercept parameter across raters, however, was not conducted in this study. Thus, strong MI was not fully examined to ensure scales are comparable across informants. In addition, as the multiple-group CFA method was used for MI testing, the data dependency between multiple raters was ignored.

More recently, some other studies also adopted multiple-group CFA and investigated full MI with multiple groups. Waschbusch and Willoughby (2008) examined the informant equivalence between mother and teacher ratings of the IOWA Conners Rating Scale on 711 elementary students in Canada. The scale includes ten items that covers two subscales: five items measured inattentive-impulsive-overactive behaviors factor and five items for oppositional defiant behaviors factor. This study employed CFA approach to investigate whether individual IOWA items were loaded to their corresponding latent factors similarly between mothers and teachers. As a result of statistically significant difference between metric and configural invariance models of these informants, a partial MI with only a constraint of equal loading of one item per factor was established. Subsequently, latent correlations were used to examine the cross-informant correspondence for the subscales in that study.

The following studies also employed CFA approach to conduct MI testing across informants. However these studies did a further step than the studies reviewed in previous section by taking into account the data dependency among scores of the raters.

Burns, de Moura, Walsh, Desmul, Silpakit, and Sommers-Flanagan (2008) tested the invariance of mother and father ratings of the Child and Adolescent Disruptive Behavior Inventory (Burns, Taylor, & Rusby, 2001a, 2001b) with 894 Brazilian, 2075 Thai and 817 American children. This study included the correlation between mother and father ratings by letting mother and father factors as well as errors of same items between these two factors be correlated.

In a following study, Burns and colleagues (2013) used mother, father, and teacher ratings with equivalent questionnaires to assess symptoms of hyperactivity, impulsivity, inattention and academic impairment of Thai and Spanish children. In this study, the Child and Adolescent Disruptive Behavior Inventory was administered to 872 Thai 7^{th} - 12^{th} graders and the ADHD Rating Scale-IV and ODD scales of the Disruptive Behavior Inventory were distributed to 1,749 Spanish children (1^{th} - 4^{th} graders). For the invariance analyses between mothers/fathers and teachers, the authors compared the baseline model with equivalent factor structure and no constraint in any parameters (i.e. configural invariance) versus the model with constraints of equal factor loadings and thresholds. While the correlations of factors of like-symptoms between parent and teacher ratings were taken into account, the dependency of students within the teachers was not considered. In addition, the residuals of same symptoms were only correlated between mother and father ratings but not for parent ratings and teacher ratings.

Clark et al. (2016) tested measurement equivalence of 93 questions that are similar across the student, teacher and parent forms of the Child Behavior Questionnaire (CBQ; Rothbart, Ahadi, Hershey, & Fisher, 2001). The MI testing procedure from configural to metric invariance was performed between mother and father ratings of 605 children in the ages of three to seven

years old in several cities of the United States. In this study, the CFA approach with inclusion of correlated errors for same items and correlated mother and father factors was also adopted.

Geiser et al. (2014) emphasized that investigation of MI across raters provide incremental information about rater effects that cannot be achieved from purely correlational multitrait-multimethod (MTMM) analyses. However the focus of literature was often on interpreting the MTMM correlation matrix and not much attention was paid for the examination of MI across raters to ensure valid interpretation of mean rater effects. The authors proposed a modeling framework to test MI across informants in the first step of CFA-MTMM analysis by evaluating the role of MI in the context of three multiple-indicator CFA-MTMM models for structurally different informants and one model for interchangeable informants. They also illustrated the model with a demonstration data of mother, father, and teacher ratings of two ADHD subscales from the Child and Adolescent Disruptive Behavior Inventory (CADBI, Burns & Lee, 2010a, b) on 709 first grade children from the island of Majorca in the Balearic Islands and Madrid (Spain). The full MI analyses were examined for both subscales across raters and partial MI was performed when a strong or strict MI did not hold. Burns et al. (2013) suggested MI testing for the model for interchangeable methods (i.e. raters) to make a decision of whether raters are truly interchangeable or not.

Lately, study of Piskernik et al. (2018) was among not many studies that used CFA approach and took into account the data dependency to examine MI between mother and father ratings. The German version of the Parenting Stress Index (Tröster, 2011) with 48 of the 101 original items was used to measure parental stress of 214 Austrian couples with their young children (12 to 32 months). MI models for both parents from configural to strict invariance were conducted simultaneously with correlated errors for similar items between mothers and fathers.

MI testing will help answer the questions whether an instrument works well similarly across different informants and valid mean comparisons of informant ratings can be made. Within this framework, researchers examine the differences in informants' perceptions, interpretations, and response styles of survey items about their children, which means investigating the underlying construct of a measure rather than the observed scores. While some applied studies attempted to investigate MI across informant ratings, few of them considered the dependency of data between mother and father or parent and teacher ratings. Moreover, no simulation study has yet examined the performance of MI models for these types of data under different research scenario.

3. Measurement Invariance

Borsboom (2006), and Meredith and Teresi (2006) emphasized that thorough psychometric analyses play an essential role for fair and equitable selection procedures. Furthermore, based on the growing interest and practices of MI testing in literature (Schmitt & Kuljanin, 2008), Kim, Kwok, and Yoon (2012) suggested examining MI as a prerequisite before using an instrument in social studies.

As defined in (Mellenbergh, 1989; Meredith, 1993), the random variable y is considered as measurement invariant as regards to selection on group variable g given the latent variable η, if and only if

$$F(y|\eta, g) = F(y|\eta) \qquad (9)$$

where $F(y|\eta, g)$ is the distribution of observed scores for variable y given g and η, and $F(y|\eta)$ is the distribution of observed scores for variable y given η. Mellenbergh (1989) emphasized that this definition implies the distribution of observed scores (or item responses) is dependent on the

values of the latent variable but not on the combination of both values of the latent variable and group variable.

MI was also defined in terms of probability in Yoon and Millsap (2007) as the conditional probability of random variable y given underlying latent variable η is not dependent of group membership (i.e. group variable g).

$$P(y|\eta, g) = P(y|\eta) \qquad (10)$$

where y is the observed variable that is used to measure the latent construct η, and g denotes group membership.

MI testing is commonly conducted to examine several measurement conditions about the extent to which an instrument is being perceived and interpreted in the similar way across different groups (often subpopulation groups), over different time points or over various methods of measurement (Vandenberg & Lance, 2000; Meade & Bauer, 2007; Kim, 2011). The traditional subpopulation groups are ethnicity, gender, and age and recently include countries or cultural groups (Kim, 2011). For longitudinal studies using the same measure over different time points, it is also essential to ensure the measure is invariant over the time. In studies where the mediums of a measurement vary such as the pencil-and-paper vs. online forms or original vs. translated versions of a test, the equivalence of this measure across these mediums is often of concern.

4. Sequence of Measurement Invariance Testing

As emphasized in Kim's dissertation (2011), the full invariance of a factor model (i.e. the equivalence of variance covariance matrices of observed variables) across groups is "not easily attainable in reality, but also it is not necessary in practice". Therefore the typical MI analysis is a hierarchical procedure including the testing of several nested invariance models from

configural, metric, scalar to strict invariance (i.e. forward procedure, Kim et al., 2011) or in the opposite order (i.e. backward procedure, Kim et al., 2011). The forward MI testing procedure includes the steps below:

- **Step 1**: the configural invariance with an identical factor structure (i.e. same number of factors and pattern of loadings) across groups of raters. In configural invariance testing, there are no equality constraints on factor loadings, intercepts or residual variances except the minimal constraints for identification purpose. Configural invariance model is considered as a baseline for following MI tests and these following MI tests can be performed only if configural invariance is achieved. The configural invariance is violated due to the absence of invariant factor structures among groups (Kim, 2011).
- **Step 2**: the metric invariance level (also called weak invariance in Meredith, 1993 or pattern invariance in Gregorich, 2006). Assuming configural invariance is satisfied, the metric invariance model is achieved if corresponding factor loadings are equal across groups. As a pattern regression coefficient, a factor loading presents relationship between an observed variable and a common factor. Equal factor loadings across groups would indicate that corresponding common factors have identical meanings among these groups. The variances covariances of common factors can be only identified by achievement of metric invariance and also are not influenced by the rescaling of latent variables once the metric variance holds (Kim, 2011). The attainment of metric invariance level is also necessary for a defensible comparison of estimated factor variances and covariances (Gregorich, 2006). However, this author also noted that group discrepancies in common factor variation and covariation might not be a reflection of group differences in observed variation and covariation when metric invariance is established but strict invariance does not hold.

- **Step 3**: the scalar invariance (or also called as strong invariance in Meredith, 1993) implies the equivalence of item intercepts across methods. Differences in intercepts often reflect "differential additive response bias", which might systematically result in the differences in valued item response between subpopulations. These sources of bias such as different cultural norms or procedural differences in taking weight measurements are not related to the common factors and have impact on observed means rather than response variation (Gregorich, 2006). Establishment of invariant factor loadings and item intercepts provides evidence for a defensible comparison of factors and observed means (Gregorich, 2006). Scalar invariance is tested with constraints of equalities of factor structure, factor loadings, and intercepts for corresponding items across methods. While CFA models of only covariance structure is sufficient to test metric invariance, it is required to fit CFA models of both covariance and mean structures to test scalar invariance.
- **Step 4**: the strict invariance refers to the equality of unique factor variances (or residual variances) across groups. This level of MI is required for "defensible comparisons of both observed mean and variance estimates across population groups" (Gregorich, 2006). However, as noted by Gregorich, in reality, researchers are more interested in group mean comparisons than observed variance estimates so it is not practical to test strict invariance. The strict invariance model includes equality constraints on factor structure, factor loadings, item intercepts, and unique factor variances for corresponding items across methods.

As highlighted by Geiser et al. (2014), MI testing for multiple informants that use equivalent questionnaires not only provides useful information on the rater effects but also ensure if comparison of latent means across informants is meaningful. They also stated that similar to comparing latent means across groups in multiple-group CFA and across different time

points in longitudinal analyses, comparisons of raters' means require a certain MI level among them. Specifically, Geiser et al. (2014) suggested at least scalar MI (i.e. equal loadings and intercepts) being held to make a meaningful comparison across raters because scalar MI guaranteed similar origins and units of measurements among different informant ratings. Other studies also mentioned that scalar invariance is a sufficient condition for a meaningful comparison between group means (e.g. Meredith, 1993; Gregorich, 2006; Kim, Cao, Wang, & Nguyen, 2017). Therefore, in this dissertation, I applied forward MI testing procedure as described above but did not include strict invariance with constraints of equal residual variances for MI testing in Study 1 and Study 2.

5. Intraclass Correlation (ICC): Item ICC and Factor ICC

The intraclass correlation coefficient (ICC) in the multilevel (also called hierarchical) modeling context is used to measure the level of statistical dependency in the data. It is generally calculated by the ratio of between variance (or group-level error variance) over total error variance and ranges from 0 to 1. For example, an ICC for a two-level hierarchical model is calculated as:

$$ICC=\sigma^2_{between}/ (\sigma^2_{between} + \sigma^2_{within})$$

where $\sigma^2_{between}$ is the variance of the level-2 residuals and σ^2_{within} is the variance of the level-1 residuals ("Multilevel Modeling Tutorial", 2015)

The ICC would be zero or near zero when there is no statistical dependency in the data, indicating total variance would come from individuals (or level 1 variable). The ICC closer to 1 implies highly dependent data where the majority of variance would be from groups (or clusters, i.e. level 2 variables).

There are two kinds of ICC often used in multilevel CFA modeling: latent factor ICC and item (or observed variable) ICC (Hsu, Lin, Kwok, Acosta, & Willson, 2017). The latent factor ICC in a multilevel CFA model is calculated as:

$$\text{Latent factor ICC} = B / (B + W)$$

where B = the latent factor variance at the between-level, W = the latent factor variance at the within-level.

Because the identical model structure assumption (i.e. equal model structures for both between - level and within - level models) is not met for the partially nested data in Study 2, latent factor ICC is not used in this study. Rather, the observed variable ICC for teacher rating items is used in Study 2 to take into consideration the data dependency between teachers and students. According to Hsu et al. (2017), item ICC is the proportion of variance of an observed variable coming from between-group variation and is calculated as the ratio of between-level variance to the total variance of that observed variable (Hsu et al., 2017; Raudenbush & Bryk, 2002; Snijders & Bosker, 2012):

$$\text{Observed variable ICC} = b / (b + w)$$

where b = (between-level factor loading)^2 x between-level factor variance + between-level residual variance, and w = (within-level factor loading)^2 x within-level factor variance + within-level residual variance.

The use of multiple informants with similar questions (e.g. identical questions across parent form, teacher form and student form) is a common design (Geiser et al., 2014). However, as described earlier in this chapter, many studies in the current literature on MI testing across multiple rating resources often do not take into account the repeated characteristics of parental ratings or the partial nesting characteristics of parent and teacher scores. Furthermore, there is no

simulation study that investigates the performance of different approaches to test MI with multiple informants at various research scenarios, especially the methods that take into account the repeated or partially nested features of the data. In this dissertation, my focus is testing MI of equivalent assessments (e.g. surveys) with similar questions across raters (e.g. mothers and fathers for inter-parental, parents and teachers for cross adult informants). Although in many psychological assessments, some items can be different across various forms, only identical questions across those forms are used to conduct MI testing. In addition, although investigating MI includes both qualitative evaluation of item meaning and quantitative investigation of invariance levels of factors and items (Meredith & Teresi, 2006), this dissertation concentrates on the quantitative aspect of MI testing, i.e. statistical assessment on an instrument. It includes two simulation studies that examine the two proposed models along with other comparative models for repeated measure and partially nested data. The details of the simulation design and answers to each research question will be described in Chapter three.

CHAPTER THREE: METHOD

This dissertation attempts to investigate the statistical performance of the two proposed models, the repeated CFA model for paired data and the multilevel repeated CFA for partially nested data, as well as the commonly used multiple-group CFA model for both types of data and the potential design-based multilevel CFA for the partially nested type. Two simulation studies are included in the dissertation: Study 1 for paired data and Study 2 for partially nested data. This chapter describes the design of each simulation study with details of simulation factors as well as the plan to analyze the simulation outcomes.

1. Simulation Design for the Two Studies

The two studies are Monte Carlo simulations with a partial crossed-factorial design. There are six simulation factors (number of items, location of measurement noninvariance, magnitude of noninvariance, magnitude of correlation between two informant scores, magnitude of correlation between two unique factors, and sample size) for Study 1 and eight factors (number of items, location of measurement noninvariance, magnitude of noninvariance, magnitude of correlation between two informant scores, magnitude of correlation between two unique factors, number of level-2 units, number of level-1 units per level-2 unit, and partial ICC for nested items) for Study 2. Details of these simulation factors will be described below. The selection of values for these simulation factors are based on data from applied studies using common emotional and behavioral instruments with multiple informants, meta-analyses about MI across adult raters as well as simulation studies about MI using multiple-group or multilevel CFA. For both Studies 1 and 2, factor loadings of all items were generated within the range of

.40 and .90, following the average of minimum factor loadings of 0.41 and the average of maximum loadings of 0.83 for level 1 and 0.47-0.94 for level 2 in Kim et al. (2016). This range for factor loadings is also similar to the simulation study of Kim et al. (2012). The intercepts of these items are simulated within the range of -2.0 and 2.0 (based on Kim et al., 2012).

Data generation and fitted models were conducted using Mplus 7. The Statistical Analysis System (SAS) package version 9.4 was used to analyze the impact of simulation factors on the outcomes as well as to call and run the fitted models with all replications for each condition in each simulation study. The number of replications for each simulation study is 1,000 to reach a maximum standard error of an observed proportion of .007, and a 95% confidence interval no larger than ± .0137 (Robey & Barcikowski, 1992).

1.1. Study 1: Data Generation and Simulation Factors for the Paired Data

A simple CFA model with one latent factor measured by a set of continuous items is used to generate data for Study 1. Two sets of scores using the same CFA model were generated to illustrate rating scores from two different informants (e.g., mother and father) of the same child or participant. As the two informants assess the same participant, the scores for each item from these two sets are manipulated to be correlated to each other. In order to run the multiple-group CFA model for the mother and father groups, the originally generated data for Study 1 was reorganized and stacked up with double numbers of children from the combined mother and father scores. Specifically, a group membership variable was created (with two levels of either mother or father) and the two sets of identical items for mother and father ratings were combined into one set. For example for the simulation condition of five items, the generated model had 10 items with two factors (one with five mother items and one with five father items) and included no group membership variable. But in the stacked up dataset there were only one factor with five

items and one group membership and each set of five items belonged to either mother or father group.

Factor variance and item variance for mother and father factors in Study 1 were fixed at 1. With this factor standardization strategy, because factor variance was 1, the factor correlation was equal to the factor covariance.

The following simulation factors were included in Study 1 to examine the adequacy of the two models:

1) Number of items: 5 as short and 10 as long. The "short" condition of five items per factor in simulation factor (1) was based on Kim, Dedrick, Cao and Ferron's (2016) meta-analytic study from which the average of minimum number of items per factor was 3.86 (range 1-10, *SD*=2.09). Furthermore, this value is realistic given the fact that at least three items per factor is required for a factor model to be identified and also a condition to have good reliability. The upper value (i.e., number of items = 10) was chosen based on the report from Kim et al. (2016) for maximum number of indicators per factor in level-1 as well as the typical values of nine or 10 items for subscales with longer number of items in many emotional and behavioral instruments (e.g., Internalizing Problems factor in the BESS Student Form, Inattention Symptoms in the DSM-IV ADHD Symptom Rating Scale, Patient Health Questionnaire scale that measures depression).

2) Magnitude of noninvariance: zero, small, large. Specifically, zero value is when there is MI (i.e., equal factor loadings and intercepts across mother and father factors). The small noninvariance condition is when differences are 0.2 for factor loadings and 0.3 for intercepts between two informant factors. The large noninvariance condition is when the differences are 0.4 for factor loadings and 0.6 for intercepts between two informant factors. For the noninvariance

conditions (small or large), only one item has a difference in loadings and/or intercepts between two informant factors for the simulation condition when total number of items is five (i.e., short) and two items have a difference in loadings and/or intercepts when the total number of items is 10 (i.e., long). The remaining items will have identical factor loadings and intercepts across the two factors.

3) Location of measurement noninvariance: there will have two levels of measurement nonequivalence including measurement noninvariance for both factor loading and intercept, and at intercept only. When measurement noninvariance is located at both factor loading and intercept, one item (for number of items =5) or two items (for number of items =10) in each informant factor will have both factor loading and intercept being different which made the contamination rate or noninvariance degree of 20%. This noninvariance level is similar to other simulation studies about MI testing such as Kim (2011), Kim et al. (2017). For the measurement noninvariance at intercept only, there is only difference in the intercept of one pair equivalent items (for number of items = 5) or two pairs of equivalent items (for number of items =10) across two factors and the factor loadings of all equivalent items are the same. The condition of zero noninvariance implies the strong MI holds with both factor loadings and intercepts for equivalent items across two informants equal.

4) Magnitude of correlation between two informant common factors (i.e., correlation between mother and father ratings): .4 as moderate and .7 as high. These values were selected from reviewing meta-analytic studies about correlations/agreements of parental ratings including Achenbach et al. (1987), Duhig et al. (2000), and Renk and Phares (2004) as well as other recent applied studies that conducted MI testing with multiple informant ratings. Specifically, the average of correlations for mother and father ratings was .6 in Achenbach and colleagues, .45 in

Renk and Phares, and that correlation in Duhig et al. (2000) was .46 for internalizing problems, .66 for externalizing problems, and .61 for total problems. The parental correlations for different subscales in some applied studies of emotional and behavioral problems ranged from .68 to .85 in Burns et al. (2008), .16 - .44 in Konold and Pianta (2007), and .36 - .93 in Waschbusch and Willoughby (2008).

5) Magnitude of unique factor (or error) correlation between identical items: 0.3 and 0.6 considering the factor loadings ranged from 0.4 to 0.8 and item variance and factor variance are fixed at 1. These error correlation values are also in line with the range of error correlation of between 0.2 and 0.65 from 22 studies reviewed in Kim et al. (2016).

6) Sample size (e.g. number of children): (100, 500, 1000). As there was no simulation study about MI with paired data and sample size was not reported in several meta-analyses about parental correlation for emotional and behavioral instruments, sample size selection for Study 1 is based on the applied studies that examine MI testing for mother and father ratings as well as simulation studies about MI testing using multiple group CFA or MIMIC modeling. Specifically, a review of applied studies that performed MI for parental ratings also found sample sizes within the selected range such as: 894 Brazilian, 2,075 Thai, and 817 American children in Burns et al. (2008); 605 children in Clark et al. (2016); 566 in Makransky and Bilenburg (2014); 337 in Mayfield et al. (2018); 214 in Piskernik et al. (2018); 562 in Konold and Pianta (2007); 711 in Waschbusch and Willoughby (2008). Furthermore, the selected sample sizes were used in Kim and Yoon's (2011) simulation study about testing MI with multiple-group CFA and item response theory approaches. The rationale for the use of small sample condition (i.e., 100) was based on reasonable results from previous simulation studies of Muthén and Asparouhov (2002) and Yoon (2008) about multiple group analysis. The similar sample sizes were also used in Kim

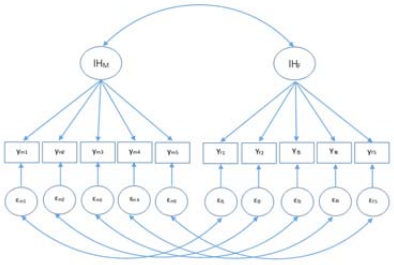
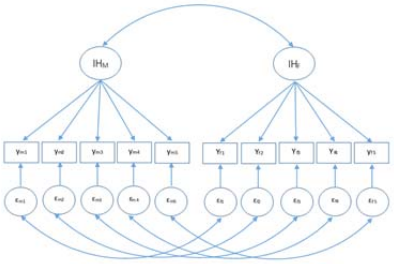
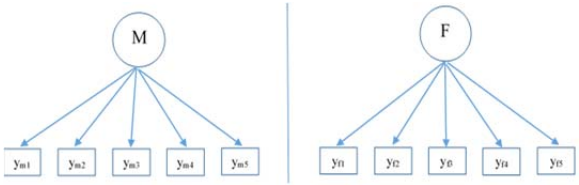
(2011) with the reason that the group sample size of 100 or larger is required to have a good power to detect measurement noninvariance in multiple-indicators multiple-causes (MIMIC) modeling (Woods, 2009).

There were 120 simulation conditions examined in Study 1.

1.2. Study 1: Fitted Models

The proposed Model 1 (the repeated measures CFA model) and Model 2 (multiple-group CFA) were fitted to the data generated in Study 1.

Table 1. Summary of 5-item models used for data generation and data analysis for Study 1

Model	Model 1	Model 2
Data generation	Single-level Repeated measure CFA 	
Data analysis	Single-level Repeated measure CFA 	Single-level multiple-group CFA 

For simplicity, Table 1 presents the simulated research scenario with five items. The generated and analysis models for the other simulated cases with ten items are similar. It should

be noted that as described in the data generation section, the stacked up dataset was used to fit the multiple-group CFA model.

1.3. Study 2: Data Generation and Simulation Factors for the Partially Nested Data

The similar CFA model to Study 1, i.e. one-factor model with a set of continuous items was used to generate data for Study 2. Additionally, in order to reflect the partially nested feature for Model 2 where one set of scores are nested within informants (e.g. scores of students nested within teachers) while the another set of scores are independent (e.g., scores of students from the parents), a multilevel modeling framework was used for this study. In this framework, ICCs for teacher items were simulated at different magnitudes (small, large) while ICCs for parent items were constrained at zero to create the partial nesting feature of Model 2. Similar to Study 1, the generated data for Study 2 were also stacked up with addition of group membership variable to indicate a certain set of items belong to parent or teacher group.

Factor variance and item variance for within-level factors in Study 2 were fixed at 1. Due to this factor standardization strategy (factor variance equals to 1), factor correlation is equal to factor covariance. The between-level item variances in the teacher factor in Study 2 vary from 0.15 to 0.5, coupling with different levels of factor loadings and residual variances to create different levels of item ICC (i.e. about 0.13, and 0.33) for the teacher items. The between-level item variances in the parent factor in Study 2 were all zeros to reflect the partial nested structure of the data in Study 2 as described in Chapter 1.

In addition to the simulation factors from (1) to (5) in Study 1, three additional factors were included in Study 2: (6) number of level-2 units (e.g., teachers), (7) number of level-1 units per level-2 unit (e.g., children per teacher), and (8) partial ICC (small, large) for the nested items. Total number of conditions simulated for Study 2 is 320. Specifically:

1) Number of items: 5 as short and 10 as long. Similar to Study 1, these values of number of items are based on Kim et al. (2016) and several existing psychological assessments such as BESS and Patient Health Questionnaire.

2) Magnitude of noninvariance: zero, small, large.

+ Zero noninvariance: equal factor loadings and intercepts across parent and teacher factors for all items.

+ Small noninvariance: 0.2 for factor loading difference and 0.3 for intercept differences. Similar to Study 1, only one item (when number of items is five) and two items (when number of items is ten) have difference in loading/or intercept between two informant factors.

+ Large noninvariance: similar to Study 1, i.e. 0.4 for factor loading difference and 0.6 for intercept differences.

3) Location of measurement noninvariance: similar to Study 1, measurement nonequivalence is located at two levels (conditions): differences in both factor loadings and intercepts, and difference at intercepts only between parent and teacher factors.

4) Magnitude of factor correlation between two informants (i.e. correlation between parent and teacher ratings): 0.3 as moderate and 0.5 as high. Selection of these levels is based on the average of correlation for parent and teacher ratings (r) from meta-analytic studies including: in Achenbach et al. (1987): r = 0.28; De Los Reyes et al. (2015): r = 0.21 for internalizing problems, 0.28 for externalizing problems; Meyer et al. (2001): r = 0.29 for summed behavioral and emotional problems. Narad and colleagues (2015) also reported the range of parent-teacher correlations on Attention Deficit/Hyperactivity Disorder (ADHD) symptoms' ratings across several studies was 0.09 - 0.43. Examples from other applied studies about emotional and

behavioral ratings include: r = 0.06 - 0.27 for father-teacher correlation and 0.10 - 0.31 for mother-teacher correlation in Konold and Pianta (2007); and r = 0.45 - 0.62 for parent-teacher correlation in Waschbusch and Willoughby (2008).

5) Magnitude of error correlation between two identical items: Given the range of selected values for factor loadings (0.4 to 0.9) as well as item variance and factor variance for the within level are fixed at 1 in the simulation, the unique factor (or error) correlation between two identical items of two factors are simulated at 0.3 and 0.6. In the multilevel CFA studies that Kim et al. (2016) reviewed, the error correlations ranged from .20 to .65 (n = 22) with an outlier (.08).

6) Number of level-1 units per level-2 unit (e.g., children per teacher): 10 and 20. These values are popular for classroom size in the United States, especially for younger children classrooms. These are also recommended in Hox (1998) as popular cluster size in multilevel research and have been used in other simulation studies about testing MI using multilevel CFA framework (e.g. Kim et al., 2012; Kim and Cao, 2015).

7) Number of level-2 units (i.e., number of teachers): 30 and 80. The cluster size or number of level-2 units is often not reported in meta-analytic as well as applied studies about cross informant ratings. Thus selection of cluster sizes is based on the simulation study of Kim et al., 2012 as well as from a few applied studies about MI testing across multiple informants that listed number of teachers completed the ratings (e.g. Waschbusch & Willoughby, 2008 with cluster size of 66 teachers; Gresham, Elliott, Cook, Vance, Kettler, 2010 with 54 teachers; and Burns et al., 2013 with 80 teachers). Combining with the sample size of level 1 condition, the total sample size for the current study will range from 300 (10*30) to 1,600 (20*80) children.

8) Partial item ICC: approximately 0.13 and 0.33 for the nested items (i.e. items in the teacher factor). These values of item ICC fall within the average values of minimum item ICC (0.13) and maximum (0.34) from multilevel CFA studies reported in the meta-analytic paper of Kim et al., 2016. They are also similar to the levels of item ICC in Im, Kim, Kwok, Yoon, & Willson (2016): 0.08-0.13 for small, 0.13-0.25 for medium, and 0.20–0.46 for large ICC.

1.4. Study 2: Fitted Models

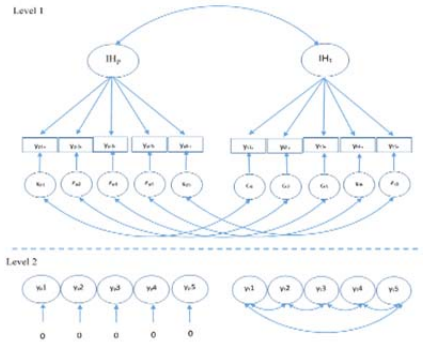
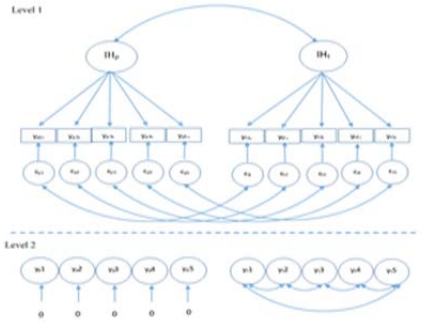
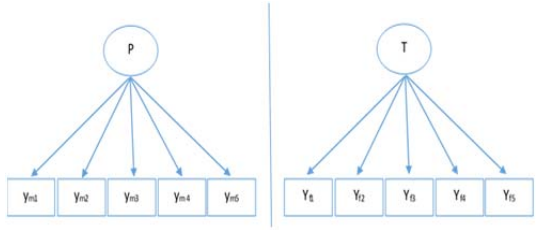
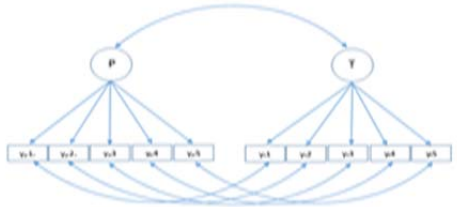
The proposed repeated measure multilevel model (Model 3), the multiple-group CFA (Model 4) and the design-based multilevel CFA model (Model 5) were used to fit to the data generated in Study 2. For the multiple-group CFA model, generated data for Study 2 were stacked up of parent and teacher ratings to run the MI testing for both groups separately. The summary of five-item models used for data generation and data analysis for this study is presented in Table 2.

2. Model Evaluation for Study 1 and Study 2:

MI testing for both simulation studies was conducted using likelihood ratio (LRT, also called Chi-square difference or $\Delta\chi 2$ test and the three terms are used interchangeably in this dissertation), CFI difference (ΔCFI,) and RMSEA difference (ΔRMSEA) tests between a baseline model and a sequentially more constrained model.

Chi-square difference test is one of the most frequently used tests to examine global model fit (Van de Schoot, Lugtig & Hox, 2012). The default estimation method for multilevel CFA in Mplus is the maximum likelihood estimation with robust standard errors (MLR). This estimation method is adjusted for data dependency and data nonnormality from complex sampling.

Table 2. Summary of 5-item models used for data generation and data analysis for Study 2

Model	Model 3	Model 4	Model 5
Data generation		Multilevel Repeated measure CFA	
			
Data analysis	Multilevel Repeated measure CFA	Single-level multiple-group CFA	Design-based multilevel CFA
			

The generated and fitted models with ten-item are similar to these five-item models.

Satorra–Bentler scaled likelihood ratio test (SB LRT) is suggested for model comparison using the MLR estimator (Satorra & Bentler, 1994). However this test was reported to produce negative values that involve additional correction (Asparouhov & Muthén, 2012; Satorra & Bentler, 2010). Both the regular likelihood ratio test and the SB LRT were used for the present simulation studies. The regular LRT was conducted with all five models but the SB LRT was only performed for the two models that used MLR estimation methods (i.e. Model 3 and Model 5).

As the only use of Chi-square difference test to evaluate model fit may result in over-rejection of measurement invariance tests when the total sample size was large (Putnick & Bornstein, 2016), CFI difference and RMSEA difference tests were also used to evaluate the performance of each model in this dissertation. The suggested cutoff values of 0.01 for ΔCFI (Cheung & Rensvold, 2002) and 0.015 for ΔRMSEA (Chen, 2007) for MI testing were employed to compare the two sequential models.

Specifically, to test metric invariance (i.e. equal factor loadings between two informant groups), a configural invariance model (i.e. a model with freely estimated factor loadings and intercepts or baseline model) was created and compared to a metric invariance model (i.e. a model with factor loadings constrained to be equal for two groups). When the null hypothesis of no difference between two models (i.e. the baseline model and the more restricted model) is failed to reject at nominal alpha of .05, the more restricted (or constrained) invariance model holds. For example, for comparison of metric vs. configural models, the p value > .05, or the $\Delta CFI \leq .010$, or $\Delta RMSEA \leq .015$ implies the metric invariance holds under study. Conversely, the less restricted invariance model (i.e. configural invariance) holds when the null hypothesis is

rejected (i.e. $p \leq .05$ or ΔCFI>0.010, or ΔRMSEA > .015), providing an evidence of (a) noninvariant factor loading (s).

If the null hypothesis of no differences between two models (metric vs. configural) is rejected, the metric invariance does not hold and there was no need to proceed to the next step, i.e. testing the scalar invariance vs. metric invariance. However if the null hypothesis was failed to reject ($p > .05$, ΔCFI $\leq .010$, or ΔRMSEA $\leq .015$) implying metric invariance was satisfied, a scalar invariance model with both factor loadings and intercepts imposed equally between two informant groups was compared to the metric variance model. If the result favors the more restricted model ($p > .05$), the scalar invariance holds under study. If the result indicates rejecting the null hypothesis ($p \leq .0.5$), the scalar invariance does not hold.

3. Answer to Research Question 1

The adequacy of each model in detecting MI will be evaluated based on the correct detection rates of the level of MI (i.e. configural, metric or scalar) for each model.

In order to compare and select a better fitting model, I consider using the likelihood ratio test and calculating the correct detection rate for each simulated condition. The correct detection rate is the proportion of cases where the level of invariance is correctly detected by the series of $\Delta\chi^2$, ΔCFI, and ΔRMSEA tests among all 1000 simulation replications. For example, when metric invariance was generated in the population and if the series of MI testing supports metric invariance, this case is considered as correct detection.

The determination of correct detection for each type of invariance model is described as following (also see Table 3 for summary):

- For configural invariance conditions (i.e. noninvariance in both intercepts and factor loadings), if the result of testing between metric vs. configural invariance models favors

the configural model (i.e. the null hypothesis of no difference between two models is rejected with $p \leq .05$ or ΔCFI>0.010, or ΔRMSEA > .015), the noninvariance is successfully detected by the $\Delta\chi^2$, ΔCFI, or ΔRMSEA test. There is no need to test the metric invariance vs. scalar invariance for these conditions because neither metric invariance nor scalar invariance holds in the population.

- For metric invariance conditions (i.e. only noninvariance in intercepts but loadings are equal between two informant factors): First the likelihood ratio test for metric invariance vs. configural invariance models is conducted. If the null hypothesis of no difference between the two models is rejected ($p \leq .05$, or ΔCFI>0.010, or ΔRMSEA > .015) then the configural invariance holds, indicating noninvariance is falsely detected. Because metric invariance is (falsely) rejected, no further test is conducted and this case is considered as incorrect detection (configural invariance is supported when metric invariance is true). If the null hypothesis is failed to reject ($p > .05$, ΔCFI $\leq .010$, or ΔRMSEA $\leq .015$), it is necessary to continue performing the likelihood ratio, ΔCFI, and ΔRMSEA tests for scalar invariance vs. metric invariance models to determine if the metric invariance really holds under study. If the result of this test shows that the metric invariance model is favored ($p \leq .05$, or ΔCFI>0.010, or ΔRMSEA > .015), the noninvariance is correctly detected (metric invariance is supported when metric invariance is true). If $p >.05$, ΔCFI $\leq .010$, or ΔRMSEA $\leq .015$ for the $\Delta\chi^2$, ΔCFI and ΔRMSEA tests between metric invariance vs. scalar invariance models, the scalar invariance holds and this case is counted as incorrect detection (scalar invariance is supported when metric invariance is true).

- For scalar invariance conditions (i.e., equalities in both intercepts and factor loadings between two informant factors): First, the likelihood ratio test for configural invariance vs. metric invariance models is carried on. If the configural invariance holds ($p \leq .05$, or ΔCFI >0.010, or ΔRMSEA >.015), this case is counted as incorrect detection without further testing. If the null hypothesis is failed to reject ($p > .05$, ΔCFI $\leq .010$, or ΔRMSEA $\leq .015$), it is necessary to continue performing the likelihood ratio test for metric invariance vs. scalar invariance models to determine if the scalar invariance holds under study. If the result of this test shows that the metric invariance model is favored ($p \leq .05$, or ΔCFI>0.010, or ΔRMSEA > .015), the detected level of MI is incorrect. If $p > .05$, ΔCFI $\leq .010$, or ΔRMSEA $\leq .015$ for the $\Delta\chi^2$, ΔCFI, and ΔRMSEA tests, respectively between metric invariance vs. scalar invariance models, the scalar invariance holds and the level of MI is correctly detected.

Table 3. Calculation of detection rate for likelihood ratio testing

Testing simulated conditions \ $\Delta\chi2$	Metric invariance vs. configural invariance		Scalar invariance vs. metric invariance	
	$p \leq .05$	$p > .05$	$p \leq .05$	$p > .05$
Configural (NI in both loadings and intercepts)	correct = 1	correct = 0	Not conducted	Not conducted
Metric (NI in intercepts only, equal loadings)	correct = 0	Not yet counted	correct = 1	correct = 0
Scalar (equal factor loadings and intercepts	correct = 0	Not yet counted	correct = 0	correct = 1

Note: p = p values of each $\Delta\chi^2$ test for two competing models, correct = correct detection rate, NI = noninvariance.

Table 4. Calculation of detection rate for CFI difference and RMSEA difference testing

ΔCFI/ΔRMSEA / Simulated conditions	Metric invariance vs. configural invariance		Scalar invariance vs. metric invariance	
	ΔCFI ≤ .01 ΔRMSEA ≤.015	ΔCFI > .01 ΔRMSEA>.015	ΔCFI ≤ .01 ΔRMSEA ≤.015	ΔCFI > .01 ΔRMSEA>.015
Configural (NI in both loadings and intercepts)	correct = 1	correct = 0	Not conducted	Not conducted
Metric (NI in intercepts only, equal loadings)	correct = 0	Not yet counted	correct = 1	correct = 0
Scalar (equal factor loadings and intercepts)	correct = 0	Not yet counted	correct = 0	correct = 1

Note: NI = noninvariance

4. Answer to Research Question 2

The relationships between simulation factors and the performance of each model using one of three criteria ($\Delta\chi^2$, ΔCFI, ΔRMSEA) for Models 1, 2 and 4 and using one of four criteria ($\Delta\chi^2$, Satorra-Bentler likelihood ration, ΔCFI, ΔRMSEA) for Models 3 and 5 were examined by calculating eta-squares from analysis of variance using SAS 9.4. The moderate effect size of .058 suggested by Cohen's (1992) was used as a cutoff value for further examination of significant factors.

CHAPTER FOUR: RESULTS

The purposes purpose of this dissertation was to examine how well the four statistical models (i.e., repeated measures CFA, multiple group CFA, multilevel repeated measures CFA, and design-based multilevel CFA) detect different levels of measurement invariance with paired and partially nested data in various research scenarios. The setting of paired data is when the two raters are rating the same participant such as assessments from mother and father on their child's anxiety. The setting for partially nested data is when the participant is singleton to one rater (e.g., parent) but nested to another rater (teacher). For example, ratings of children depression from parents and psychiatrist and in this scenario, the multiple patients are nested within one psychiatrist but each patient is singleton to their parents. While multiple group CFA has been a commonly used approach in the literature to test measurement invariance with both types of data, the other three models have been rarely used but proposed in this dissertation to test measurement invariance with these two different kinds of data.

Two simulation studies were conducted to investigate the performance of these statistical models with the two data types: repeated measures CFA (Model 1) and multiple group CFA (Model 2) in Study 1, and multilevel repeated measures CFA (Model 3), multiple group CFA (Model 4), and design-based multilevel CFA (Model 5) in Study 2. Results from the two simulation studies are presented in this chapter. Because the detection rates of all models across configural, metric and scalar invariance were relatively similar between two levels of factor correlation in Study 2, only conditions with small factor correlation are presented for this study for simplicity. Only one and 11 cases out of 1000 replications in two conditions where the

multilevel repeated measure CFA model did not converge and there was no problem of convergence in all other simulation conditions examined in this dissertation of other models.

1. Results of Study 1

There were six simulation factors investigated in Study 1: 1) number of items (5 and 10); 2) noninvariance location (noninvariance in both factor loadings and intercepts, and noninvariance only in intercepts); 3) magnitude of noninvariance (zero, small, large); 4) factor correlation (0.4 and 0.7); 5) sample size (100, 500, 1000 participants); and 6) error correlation (0.3 and 0.6). Results of Study 1 are shown in the following section.

1.1. Detection rates of Model 1 and Model 2

This section presents detection rates of Model 1 and Model 2 using Chi-square difference test, CFI difference with cut-off value of 0.01 and RMSEA difference with cut-off value of 0.015 for configural, metric and scalar invariance conditions.

1.1.1. Detection Rates for Configural Invariance Conditions

Table 5 presents detection rates of Models 1 and 2 using $\Delta\chi 2$ test, suggested cut-off values of 0.01 for ΔCFI and 0.015 for ΔRMSEA for configural invariance with 5-item conditions (i.e. conditions where there was noninvariance in both intercepts and factor loadings and the number of items per factor was five). The results from this table show that when magnitude of noninvariance is big (i.e., 0.6 difference in intercepts and 0.4 difference in factor loadings) together with large sample sizes (i.e. 500 and 1000 participants), both two models for Study 1 could detect the noninvariance using all three criteria of $\Delta\chi 2$ test but the detection rates were consistently higher for Model 1 (89% to 100%) than those of Model 2 (79%-100%). With large noninvariance but small sample size (100) conditions, while the detection rates of Model 1 decreased to the range of 74% - 91% depending on levels of error correlations (higher error

correlation associated with higher detection rates), those rates of Model 2 were ranged between 55% and 61% (higher error correlation associated with lower detection rates because error correlations were ignored in Model 2).

Table 5. Detection rates of Models 1 and 2 for configural invariance with 5-item conditions

Detection rate using $\Delta\chi 2$ test		Detection rate using ΔCFI		Detection rate using ΔRMSEA		# of Items	NI size	Factor Corr	Sample size	Error Corr
M1	M2	M1	M2	M1	M2					
0.26	0.18	0.28	0.25	0.27	0.30	5	Small	0.4	100	0.3
0.36	0.13	0.31	0.19	0.36	0.24	5	Small	0.4	100	0.6
0.89	0.80	0.26	0.23	0.60	0.73	5	Small	0.4	500	0.3
0.98	0.80	0.30	0.19	0.83	0.71	5	Small	0.4	500	0.6
1.00	0.99	0.25	0.22	0.83	0.91	5	Small	0.4	1000	0.3
1.00	0.99	0.33	0.17	0.97	0.92	5	Small	0.4	1000	0.6
0.27	0.16	0.28	0.23	0.30	0.27	5	Small	0.7	100	0.3
0.43	0.10	0.34	0.16	0.42	0.21	5	Small	0.7	100	0.6
0.92	0.80	0.27	0.22	0.68	0.71	5	Small	0.7	500	0.3
0.99	0.79	0.37	0.15	0.90	0.69	5	Small	0.7	500	0.6
1.00	0.99	0.28	0.22	0.89	0.91	5	Small	0.7	1000	0.3
1.00	1.00	0.43	0.14	0.99	0.93	5	Small	0.7	1000	0.6
0.70	0.61	0.72	0.70	0.68	0.70	5	Large	0.4	100	0.3
0.86	0.58	0.80	0.68	0.82	0.69	5	Large	0.4	100	0.6
1.00	1.00	0.99	0.99	1.00	1.00	5	Large	0.4	500	0.3
1.00	1.00	1.00	0.99	1.00	1.00	5	Large	0.4	500	0.6
1.00	1.00	1.00	1.00	1.00	1.00	5	Large	0.4	1000	0.3
1.00	1.00	1.00	1.00	1.00	1.00	5	Large	0.4	1000	0.6
0.74	0.59	0.74	0.70	0.71	0.69	5	Large	0.7	100	0.3
0.91	0.55	0.84	0.66	0.87	0.68	5	Large	0.7	100	0.6
1.00	1.00	0.99	0.99	1.00	1.00	5	Large	0.7	500	0.3
1.00	1.00	1.00	1.00	1.00	1.00	5	Large	0.7	500	0.6
1.00	1.00	1.00	1.00	1.00	1.00	5	Large	0.7	1000	0.3
1.00	1.00	1.00	1.00	1.00	1.00	5	Large	0.7	1000	0.6

Note: M1 = Model 1, M2 = Model 2, NI size =magnitude of noninvariance, Factor Corr = factor correlation, Error Corr = error correlation

For simulation conditions of five items per factor in addition to a small magnitude of noninvariance (i.e., 0.3 difference in intercepts and 0.2 difference in factor loadings), Model 1 could sometimes (16% to 43%) detect the noninvariance (i.e., configural invariance) with small sample size (100) using any of three criteria but was able to detect noninvariance most of the

time (89%-100%) for large sample sizes (500, 1000) using Δχ2 test or often (60 - 97%) using ΔRMSEA. In addition, with these conditions while Δχ2 test and ΔRMSEA had high detection rates (80% - 100% and 69% - 93%, respectively) for Model 2 and (89% - 100% and 60% - 99%, respectively) for Model 1 with large sample sizes (500 and 1000), the detection rates using ΔCFI were always less than 25% for Model 2 and 43% for Model 1 even with large sample sizes.

For those four conditions of 5-item and small noninvariance + small sample size with low detection rates, when the configural invariance was not correctly detected, the metric invariance model was selected more often than scalar invariance model using all three criteria (see Table 6).

Table 6. Incorrectly detected rates of configural invariance for 5-item and small noninvariance +small sample size conditions

Model	Proportion of replications when metric invariance was incorrectly selected			Proportion of replications when scalar invariance was incorrectly selected		
	Δχ2 test	ΔCFI	ΔRMSEA	Δχ2 test	ΔCFI	ΔRMSEA
Model 1	53% - 60%	53% - 62%	48% - 56%	4% - 17%	7% - 19%	8% - 23%
Model 2	44% - 48%	49% - 51%	46% - 49%	35% - 45%	26% - 33%	24% - 32%

When the number of items per factor was increased to 10, the ability to detect configural invariance of both models also goes up if using $\Delta\chi^2$ test but does not change much (Model 1) or decreases (Model 2) if using ΔCFI or ΔRMSEA (see Table 7).

While both Models 1 and 2 were able to detect configural invariance all the time across different degrees of factor correlation, error correlation, invariance level or sample size if using the Δχ2 test, the two models could perfectly perform this task only with large noninvariance together with big sample size (i.e., 500 and 1000) if using ΔCFI or ΔRMSEA criteria. When the magnitude of noninvariance was small, the detection rates for configural invariance with ten items were always below 45% for Model 1 or less than 19% for Model 2 if using ΔCFI but were higher if using ΔRMSEA and could reach to the range of 73% to 96% for Model 1 and 61% - 88% for Model 2 with large sample size (i.e. 500, 1000) coupled with big error correlations (i.e.,

0.6) conditions. Note that the detection rates of Model 2 was slightly negatively associated with the magnitude of error correlations. When the number of items was ten per factor and the magnitude of noninvariance was large, the ability to detect configural invariance of the two models was pretty high for even small sample sizes and perfect with large sample sizes using any of the three criteria, $\Delta\chi 2$ test, ΔRMSEA, or ΔCFI.

Table 7. Detection rates of models 1, 2 for configural invariance with 10-item conditions

Detection rate using $\Delta\chi 2$ test		Detection rate using ΔCFI		Detection rate using ΔRMSEA		# of Items	NI size	Factor Corr	Sample size	Error Corr
M1	M2	M1	M2	M1	M2					
0.43	0.28	0.31	0.19	0.12	0.22	10	Small	0.4	100	0.3
0.65	0.21	0.37	0.14	0.22	0.16	10	Small	0.4	100	0.6
0.99	0.99	0.19	0.11	0.38	0.65	10	Small	0.4	500	0.3
1.00	0.99	0.30	0.08	0.73	0.63	10	Small	0.4	500	0.6
1.00	1.00	0.11	0.06	0.56	0.88	10	Small	0.4	1000	0.3
1.00	1.00	0.26	0.03	0.90	0.86	10	Small	0.4	1000	0.6
0.47	0.25	0.33	0.16	0.13	0.19	10	Small	0.7	100	0.3
0.70	0.12	0.44	0.08	0.25	0.11	10	Small	0.7	100	0.6
1.00	0.99	0.20	0.10	0.45	0.64	10	Small	0.7	500	0.3
1.00	0.99	0.41	0.05	0.80	0.61	10	Small	0.7	500	0.6
1.00	1.00	0.14	0.05	0.62	0.87	10	Small	0.7	1000	0.3
1.00	1.00	0.43	0.03	0.96	0.86	10	Small	0.7	1000	0.6
0.97	0.92	0.91	0.81	0.58	0.77	10	Large	0.4	100	0.3
0.99	0.92	0.95	0.80	0.82	0.74	10	Large	0.4	100	0.6
1.00	1.00	1.00	1.00	1.00	1.00	10	Large	0.4	500	0.3
1.00	1.00	1.00	1.00	1.00	1.00	10	Large	0.4	500	0.6
1.00	1.00	1.00	1.00	1.00	1.00	10	Large	0.4	1000	0.3
1.00	1.00	1.00	1.00	1.00	1.00	10	Large	0.4	1000	0.6
0.96	0.91	0.90	0.80	0.59	0.75	10	Large	0.7	100	0.3
0.99	0.89	0.97	0.78	0.86	0.71	10	Large	0.7	100	0.6
1.00	1.00	1.00	1.00	1.00	1.00	10	Large	0.7	500	0.3
1.00	1.00	1.00	1.00	1.00	1.00	10	Large	0.7	500	0.6
1.00	1.00	1.00	1.00	1.00	1.00	10	Large	0.7	1000	0.3
1.00	1.00	1.00	1.00	1.00	1.00	10	Large	0.7	1000	0.6

Note: M1 = Model 1, M2 = Model 2, NI size =magnitude of noninvariance, Factor Corr = factor correlation, Error Corr = error correlation

1.1.2. Detection Rates for Metric Invariance Conditions

When there was noninvariance only in intercepts but factor loadings were generated equally between two informant factors (i.e., metric invariance conditions), detection rates of metric invariance conditions using $\Delta\chi 2$ test, ΔCFI or ΔRMSEA were always high or perfect (75% or above for Model 1 and 71% or above for Model 2) across all conditions (Tables 8 and 9) and these rates were higher for 10-item conditions than those in 5-item conditions. While detection rates of Model 1 were much higher than detection rates of Model 2 when sample size was small coupled with small degree of noninvariance, those rates of Model 1 were slightly lower or similar to Model 2 in other 5-item metric invariance conditions. The slight negative impact of error correlations on the detection rates of Model 2 was not observed in the metric invariance conditions possibly because the noninvariance was present only in the mean structure (intercepts) and the misspecification in the covariance structure by omitting error correlations in Model 2 seemed to have no notable impact on the detection of noninvariance in the mean structure.

Table 9 shows that the detection rates for metric invariance with 10 items per factor were similar and nearly perfect or perfect (95% - 100%) across all these conditions if using $\Delta\chi 2$ test but slightly higher if using ΔCFI or ΔRMSEA (98% - 100%). The detection rates of Model 2 were also marginally higher than those of Model 1 for 10-item metric invariance if using $\Delta\chi 2$ test but similar if using the other two criteria. It should be noted that the detection rates of metric invariance (i.e., intercept noninvariance) were generally higher than those of configural invariance (i.e., factor loading and intercept noninvariance) regardless of the models used for invariance testing. It is well documented that factor loading noninvariance is more difficult to detect. Thus, as described in the previous section, when configural vs. metric invariance models

were compared, the metric invariance model could be falsely selected when configural invariance was true. However when metric invariance was generated, among the four conditions of 5-item + small noninvariance + small sample size where the detection rates of Model 2 were lower (65% - 73%) using three criteria, scalar invariance model was selected more often than configural invariance models for all of these criteria.

Table 8. Detection rates of models 1 and 2 for metric invariance with 5-item conditions

Detection rate using Δχ2 test		Detection rate using ΔCFI		Detection rate using ΔRMSEA		# of Items	NI size	Factor Corr	Sample size	Error Corr
M1	M2	M	M2	M1	M2					
0.83	0.70	0.77	0.67	0.75	0.70	5	Small	0.4	100	0.3
0.92	0.70	0.90	0.66	0.87	0.71	5	Small	0.4	100	0.6
0.94	0.97	1.00	1.00	0.99	0.96	5	Small	0.4	500	0.3
0.94	0.98	1.00	1.00	0.98	0.98	5	Small	0.4	500	0.6
0.95	0.96	1.00	1.00	1.00	0.98	5	Small	0.4	1000	0.3
0.95	0.98	1.00	1.00	1.00	0.99	5	Small	0.4	1000	0.6
0.83	0.71	0.77	0.68	0.75	0.73	5	Small	0.7	100	0.3
0.92	0.71	0.90	0.65	0.87	0.73	5	Small	0.7	100	0.6
0.95	0.98	1.00	1.00	0.99	0.97	5	Small	0.7	500	0.3
0.95	0.99	1.00	1.00	0.99	0.99	5	Small	0.7	500	0.6
0.96	0.98	1.00	1.00	1.00	0.99	5	Small	0.7	1000	0.3
0.95	0.99	1.00	1.00	1.00	1.00	5	Small	0.7	1000	0.6
0.93	0.96	0.92	0.93	0.91	0.88	5	Large	0.4	100	0.3
0.93	0.98	0.95	0.97	0.91	0.94	5	Large	0.4	100	0.6
0.94	0.97	1.00	1.00	0.99	0.96	5	Large	0.4	500	0.3
0.94	0.98	1.00	1.00	0.98	0.98	5	Large	0.4	500	0.6
0.95	0.96	1.00	1.00	1.00	0.98	5	Large	0.4	1000	0.3
0.95	0.98	1.00	1.00	1.00	0.99	5	Large	0.4	1000	0.6
0.93	0.96	0.93	0.94	0.91	0.91	5	Large	0.7	100	0.3
0.93	1.00	0.95	0.98	0.91	0.97	5	Large	0.7	100	0.6
0.95	0.98	1.00	1.00	0.99	0.97	5	Large	0.7	500	0.3
0.95	0.99	1.00	1.00	0.99	0.99	5	Large	0.7	500	0.6
0.96	0.98	1.00	1.00	1.00	0.99	5	Large	0.7	1000	0.3
0.95	0.99	1.00	1.00	1.00	1.00	5	Large	0.7	1000	0.6

Note: M1 = Model 1, M2 = Model 2, NI size =magnitude of noninvariance, Factor Corr = factor correlation, Error Corr = error correlation

Table 9. Detection rates of models 1and 2 for metric invariance with 10-item conditions

Detection rate using Δχ2 test		Detection rate using ΔCFI		Detection rate using ΔRMSEA		# of Items	NI size	Factor Corr	Sample size	Error Corr
M1	M2	M	M2	M1	M2					
0.95	0.98	0.98	0.99	0.99	0.98	10	Small	0.4	100	0.3
0.95	0.99	0.99	1.00	0.99	0.99	10	Small	0.4	100	0.6
0.95	0.97	1.00	1.00	1.00	1.00	10	Small	0.4	500	0.3
0.95	0.99	1.00	1.00	1.00	1.00	10	Small	0.4	500	0.6
0.95	0.98	1.00	1.00	1.00	1.00	10	Small	0.4	1000	0.3
0.96	0.99	1.00	1.00	1.00	1.00	10	Small	0.4	1000	0.6
0.95	0.99	0.98	0.99	0.99	0.99	10	Small	0.7	100	0.3
0.95	1.00	0.99	1.00	0.99	1.00	10	Small	0.7	100	0.6
0.96	0.99	1.00	1.00	1.00	1.00	10	Small	0.7	500	0.3
0.96	1.00	1.00	1.00	1.00	1.00	10	Small	0.7	500	0.6
0.96	0.99	1.00	1.00	1.00	1.00	10	Small	0.7	1000	0.3
0.95	1.00	1.00	1.00	1.00	1.00	10	Small	0.7	1000	0.6
0.95	0.98	0.98	0.99	0.99	0.98	10	Large	0.4	100	0.3
0.95	0.99	0.99	1.00	0.99	0.99	10	Large	0.4	100	0.6
0.95	0.97	1.00	1.00	1.00	1.00	10	Large	0.4	500	0.3
0.95	0.99	1.00	1.00	1.00	1.00	10	Large	0.4	500	0.6
0.95	0.98	1.00	1.00	1.00	1.00	10	Large	0.4	1000	0.3
0.96	0.99	1.00	1.00	1.00	1.00	10	Large	0.4	1000	0.6
0.95	0.99	0.98	0.99	0.99	0.99	10	Large	0.7	100	0.3
0.95	1.00	0.99	1.00	0.99	1.00	10	Large	0.7	100	0.6
0.96	0.99	1.00	1.00	1.00	1.00	10	Large	0.7	500	0.3
0.96	1.00	1.00	1.00	1.00	1.00	10	Large	0.7	500	0.6
0.96	0.99	1.00	1.00	1.00	1.00	10	Large	0.7	1000	0.3
0.95	1.00	1.00	1.00	1.00	1.00	10	Large	0.7	1000	0.6

Note: M1 = Model 1, M2 = Model 2, NI size =magnitude of noninvariance, Factor Corr = factor correlation, Error Corr = error correlation

1.1.3. Detection Rates for Scalar invariance conditions

Table 10 presents detection rates of Models 1 and 2 using Δχ2 test, ΔCFI and ΔRMSEA for scalar invariance (i.e., conditions where magnitude of noninvariance was zero and both factor loadings and intercepts were invariant between mother and father factors). Both Model 1 and Model 2 were able to detect the scalar invariance from 84% to 100% of cases, using any of the three criteria. The detection rates for each of both models were similar across different factor

correlations, sample sizes and error correlations if using Δχ2 test but they were higher for larger sample sizes (500 and 1000) if using ΔCFI or ΔRMSEA than smaller sample size (100). The detection rates of Model 2 were relatively higher than those of Model 1 if using Δχ2 test (all cases) or ΔCFI test (small sample size) but those rates of the two models were similar if using ΔRMSEA test.

Table 10. Detection rates of models 1 and 2 for scalar invariance conditions

Detection rate using Δχ2 test		Detection rate using ΔCFI		Detection rate using ΔRMSEA		# of Items	NI size	Factor Corr	Sample size	Error Corr
M1	M2	M1	M2	M1	M2					
0.89	0.95	0.86	0.91	0.84	0.85	5	Zero	0.4	100	0.3
0.88	0.98	0.91	0.97	0.85	0.94	5	Zero	0.4	100	0.6
0.89	0.96	1.00	1.00	0.98	0.95	5	Zero	0.4	500	0.3
0.90	0.98	1.00	1.00	0.97	0.98	5	Zero	0.4	500	0.6
0.89	0.94	1.00	1.00	0.99	0.97	5	Zero	0.4	1000	0.3
0.89	0.98	1.00	1.00	0.99	0.99	5	Zero	0.4	1000	0.6
0.89	0.95	0.87	0.93	0.85	0.88	5	Zero	0.7	100	0.3
0.88	1.00	0.92	0.98	0.84	0.96	5	Zero	0.7	100	0.6
0.90	0.97	1.00	1.00	0.98	0.96	5	Zero	0.7	500	0.3
0.91	0.99	1.00	1.00	0.98	0.99	5	Zero	0.7	500	0.6
0.90	0.96	1.00	1.00	0.99	0.98	5	Zero	0.7	1000	0.3
0.89	0.99	1.00	1.00	0.99	1.00	5	Zero	0.7	1000	0.6
0.90	0.97	0.95	0.99	0.98	0.97	10	Zero	0.4	100	0.3
0.89	0.99	0.98	1.00	0.97	0.99	10	Zero	0.4	100	0.6
0.89	0.97	1.00	1.00	1.00	1.00	10	Zero	0.4	500	0.3
0.90	0.99	1.00	1.00	1.00	1.00	10	Zero	0.4	500	0.6
0.90	0.98	1.00	1.00	1.00	1.00	10	Zero	0.4	1000	0.3
0.91	0.99	1.00	1.00	1.00	1.00	10	Zero	0.4	1000	0.6
0.89	0.98	0.96	0.99	0.97	0.98	10	Zero	0.7	100	0.3
0.88	1.00	0.98	1.00	0.98	1.00	10	Zero	0.7	100	0.6
0.90	0.98	1.00	1.00	1.00	1.00	10	Zero	0.7	500	0.3
0.91	1.00	1.00	1.00	1.00	1.00	10	Zero	0.7	500	0.6
0.91	0.99	1.00	1.00	1.00	1.00	10	Zero	0.7	1000	0.3
0.90	1.00	1.00	1.00	1.00	1.00	10	Zero	0.7	1000	0.6

Note: M1 = Model 1, M2 = Model 2, NI size =magnitude of noninvariance, Factor Corr = factor correlation, Error Corr = error correlation

1.2. Impact of Simulation Factors on the Detection Rates for Model 1 and Model 2

There were six simulation factors in Study 1: 1) number of items per factor (5 and 10); 2) magnitude of noninvariance (small and large); 3) location of noninvariance (noninvariance in both factor loadings and intercepts, noninvariance in intercepts only); 4) factor correlation (0.4 and 0.7); 5) sample size (100, 500, and 1000); and 6) error correlation (0.3 and 0.6). The effects of these simulation factors in addition to the type of model as well as interactions of type of model and each of the simulation factor were calculated for each of the three outcomes (i.e., criteria): chi-squared difference test, CFI difference and RMSEA difference for each of the three levels of invariance (i.e., configural, metric, and scalar) conditions for the two models in Study 1. The cut-off value for significant effect is 0.058 based on guidelines from Cohen (1992).

1.2.1. Effect Sizes for Configural Invariance Conditions

As seen in Table 11, the two simulation factors, magnitude of noninvariance and sample size had significant effect on most or all of the three criteria. While the magnitude of noninvariance had important impact on every single outcome criterion with strongest influence (η^2=0.913) on ΔCFI and least strong effect (η^2=0.130) on $\Delta\chi2$ test, sample size had important effect on the ΔRMSEA (η^2= 0.5) and $\Delta\chi2$ tests (η^2=0.549) but not significant on the ΔCFI test.

Table 11. Effect sizes of significant factors on detection rates of models 1 and 2 for configural invariance

	Sample size	*Magnitude of noninvariance*
ΔRMSEA	0.500	0.337
ΔCFI		0.913
$\Delta\chi2$ test	0.549	0.130

As seen in Figures 5 through 7, the detection rates of configural invariance conditions were much higher for larger noninvariance than smaller noninvariance. In the same manner, while the detection rates for larger sample sizes (500 and 1000) were always (if using $\Delta\chi2$) or on

average (if using ΔRMSEA) higher than 80%, the detection rates for smaller sample size (100) could drop to 10% if using ΔRMSEA or on average of about 50% if using $\Delta\chi 2$.

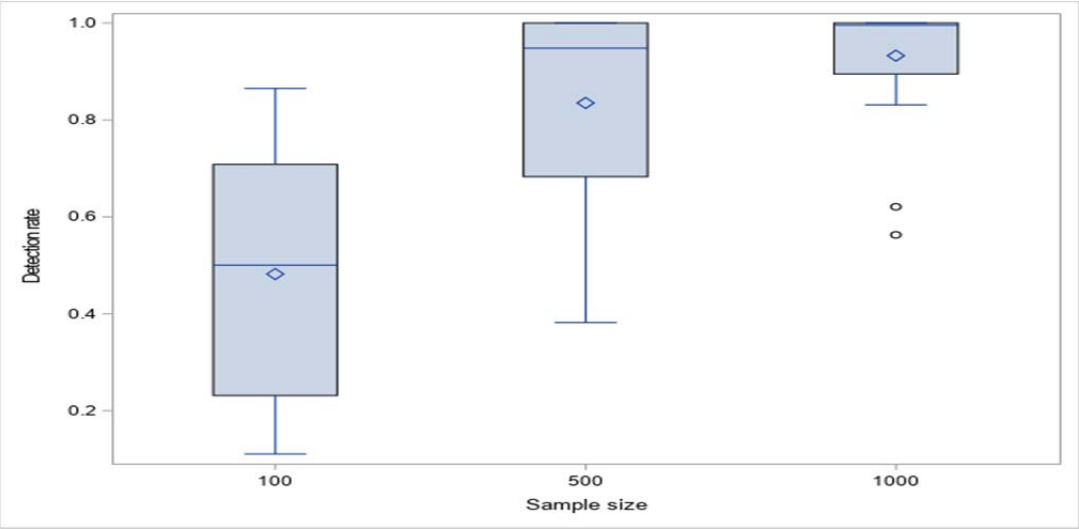


Figure 4. Distributions of detection rates of Models 1 and 2 using ΔRMSEA for configural invariance conditions by sample size

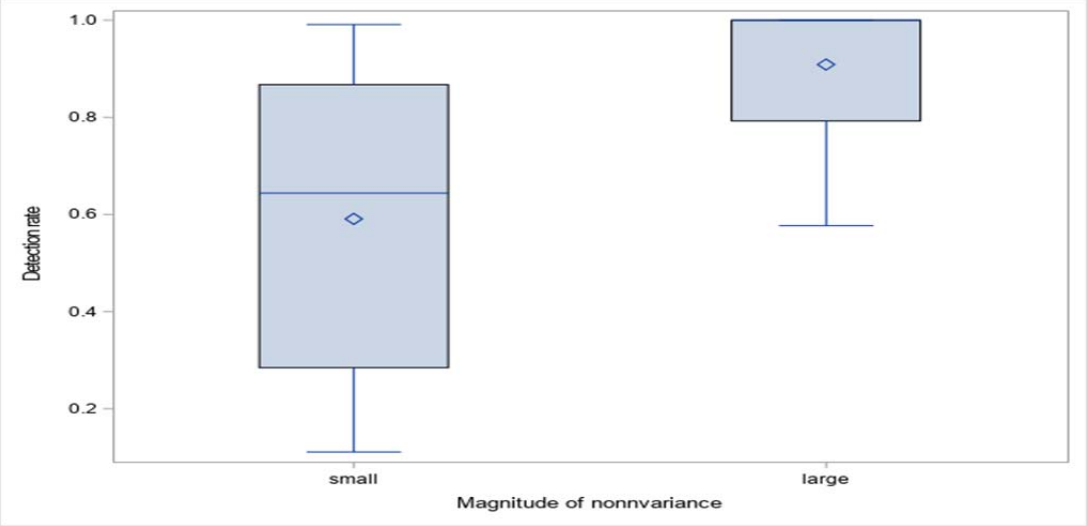


Figure 5. Distributions of detection rates of Models 1 and 2 using ΔRMSEA for configural invariance conditions by magnitude of noninvariance

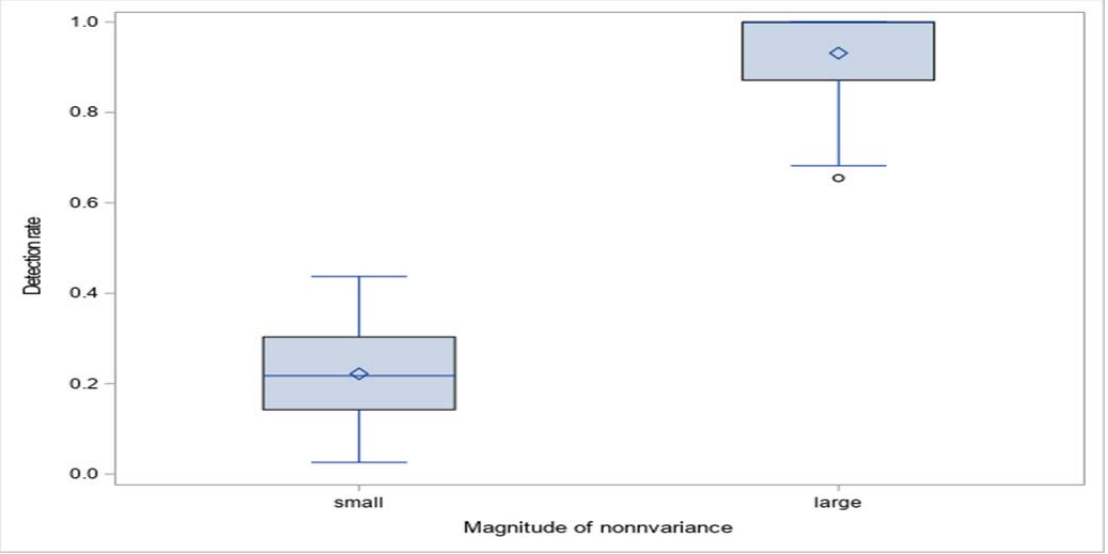


Figure 6. Distributions of detection rates of Models 1 and 2 using ΔCFI for configural invariance conditions by magnitude of noninvariance

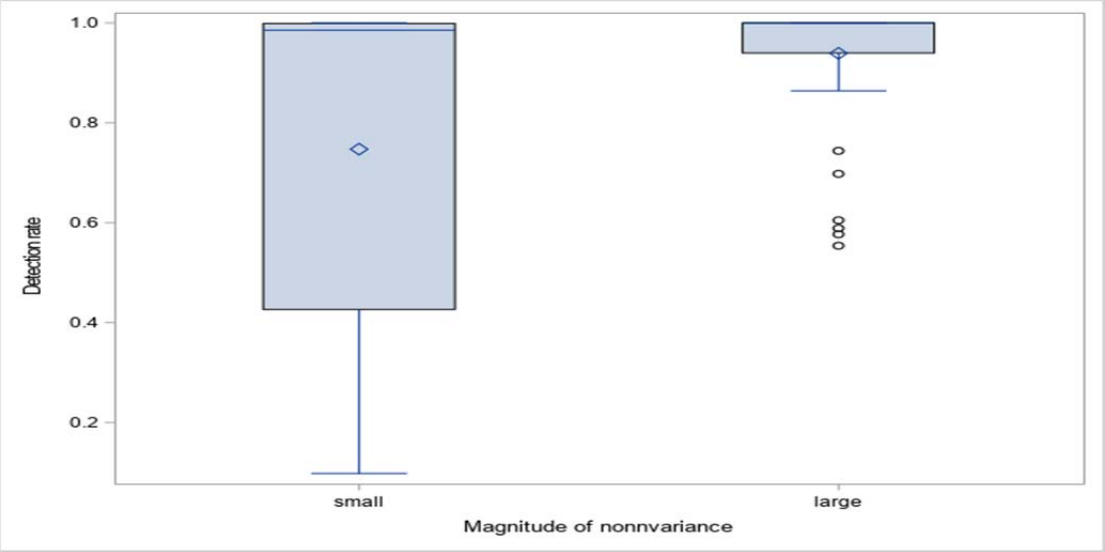


Figure 7. Distributions of detection rates of Models 1 and 2 using Δχ2 for configural invariance conditions by magnitude of noninvariance

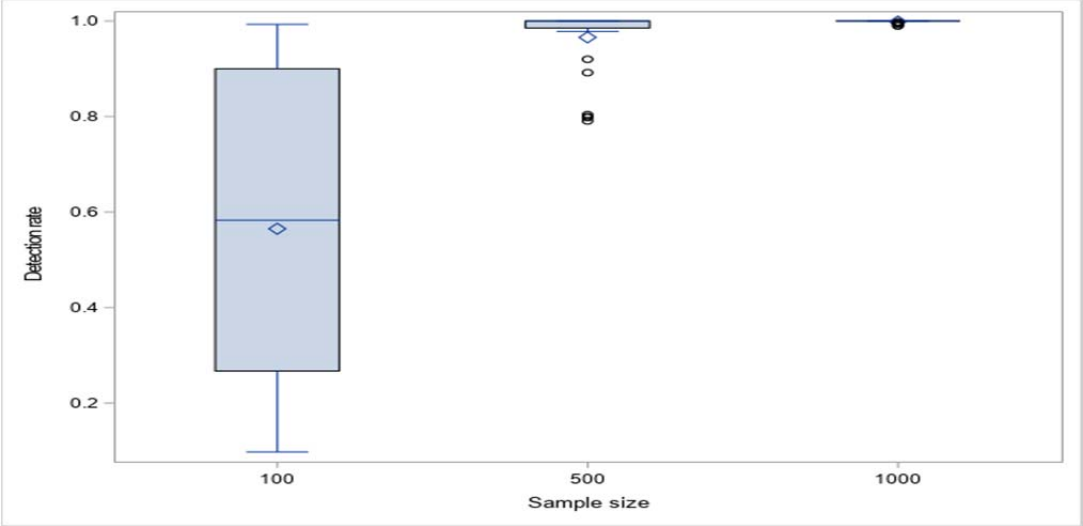


Figure 8. Distributions of detection rates of Models 1 and 2 using Δχ2 for configural invariance conditions by sample size

1.2.2. Effect Sizes for Metric Invariance Conditions

The simulation factors that have significant effect on the criteria used to measure detection rates for metric invariance conditions are shown in Table 12. Both sample size and number of items significantly impacted on the detection rates using any of three criteria with stronger effect of sample size than effect of number of items for all of these criteria.

Table 12. Effect sizes of significant factors on detection rates of Models 1 and 2 for metric invariance

	Number of items	Sample size
ΔRMSEA	0.189	0.298
ΔCFI	0.101	0.258
Δχ2 test	0.105	0.150

As seen in Figures 9 to 14, larger number of items (10) and larger sample sizes (500 and 1000) resulted in higher detection rates for all metric invariance conditions. Specifically, when the number of items was ten, the detection rates were always 95% -100% for both models using any of the three criteria regardless of sample size or other simulation factor variation. Similarly, the detection rates were also 94% or higher when sample size was 500 or 1000 even with small number of items.

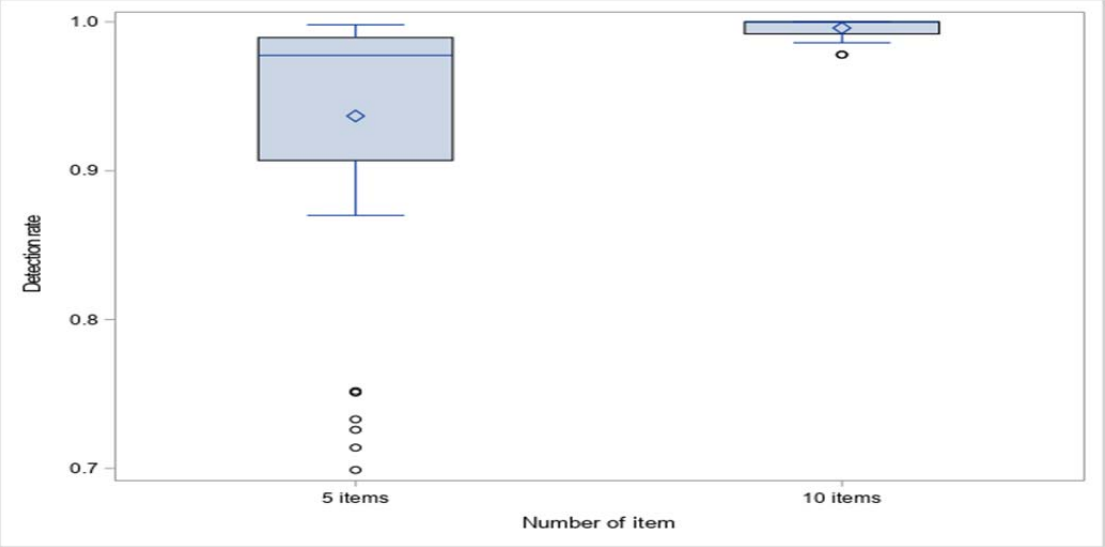


Figure 9. Distributions of detection rates of Models 1 and 2 using ΔRMSEA for metric invariance conditions by number of items

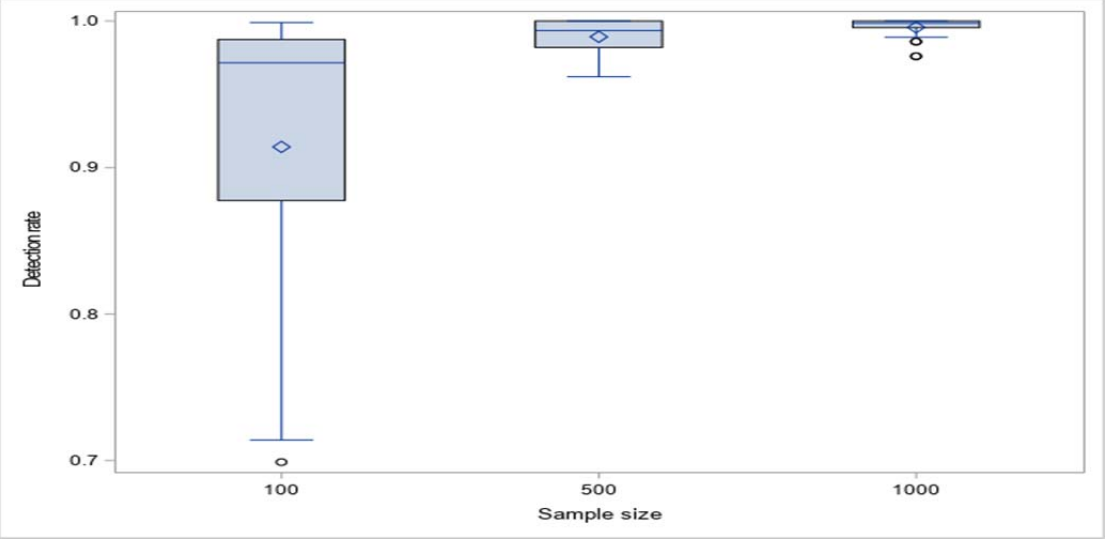


Figure 10. Distributions of detection rates of Models 1 and 2 using ΔRMSEA for metric invariance conditions by sample size

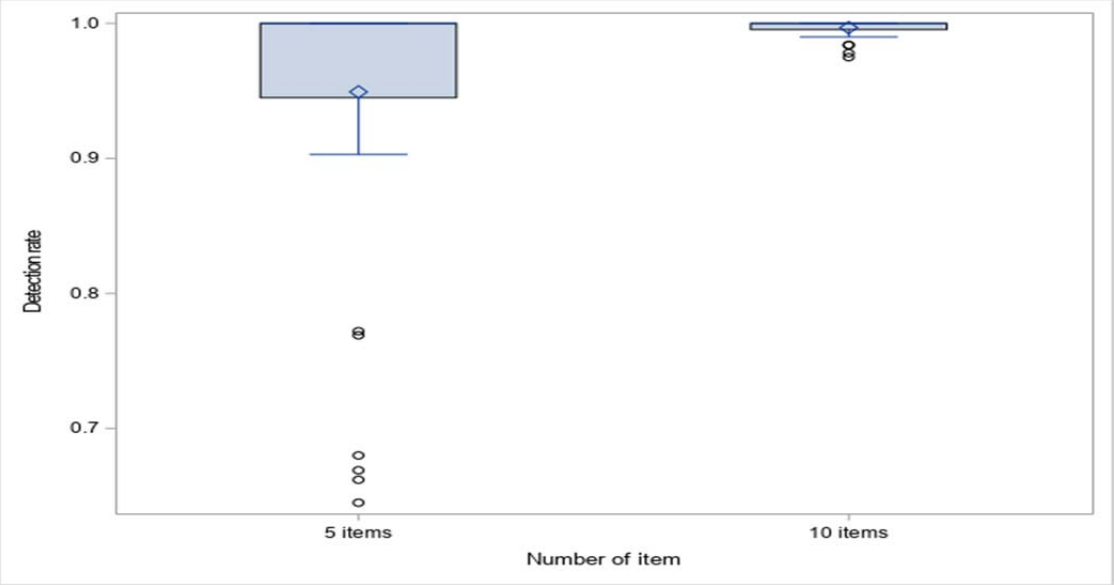


Figure 11. Distributions of detection rates of Models 1 and 2 using ΔCFI for metric invariance conditions by number of items

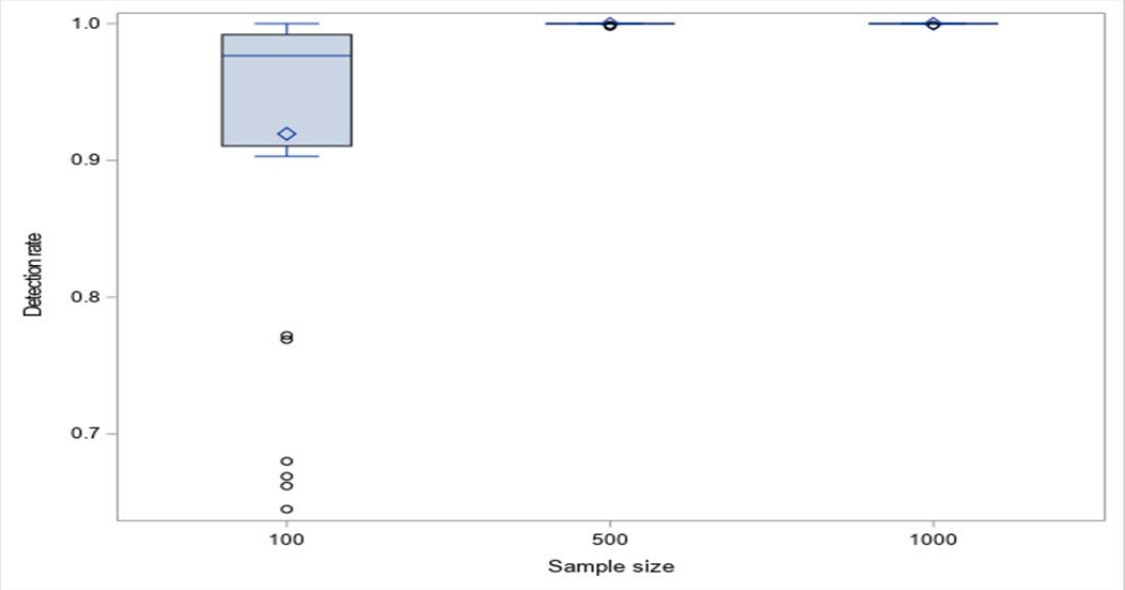


Figure 12. Distributions of detection rates of Models 1 and 2 using ΔCFI for metric invariance conditions by sample size

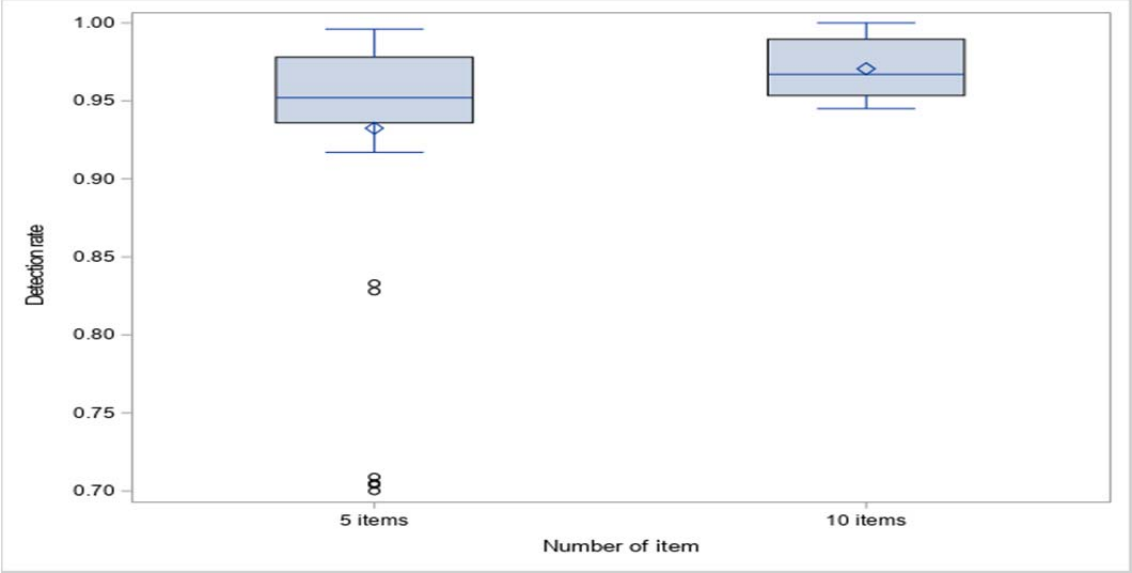


Figure 13. Distributions of detection rates of Models 1 and 2 using Δχ2 test for metric invariance conditions by number of items

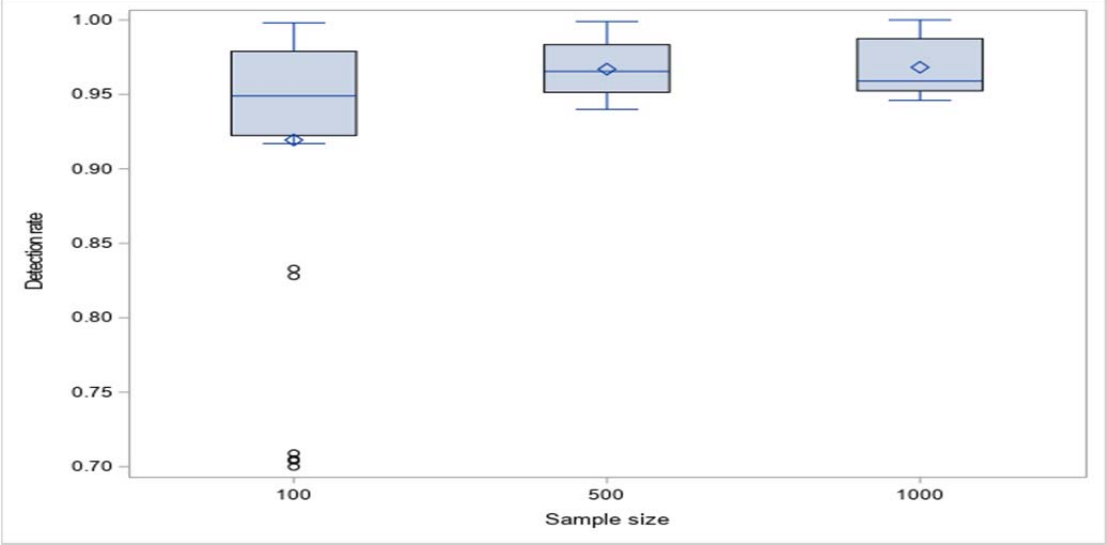


Figure 14. Distributions of detection rates of Models 1 and 2 using Δχ2 test for metric invariance conditions by sample size

1.2.3. Effect Sizes for Scalar Invariance Conditions

Eta-squared values of significant factors on detection rates of Model 1 and Model 2 for scalar invariance are presented in Table 13.

Table 13. Effect sizes of significant factors on detection rates of Models 1 and 2 for scalar invariance

	Model	Sample Size	Model * Sample size	# of items
ΔRMSEA		0.393		0.252
ΔCFI		0.470	0.076	0.093
Δχ2 test	0.910			

For scalar invariance conditions, the detection rates using both ΔRMSEA and ΔCFI received significant effect from simulation factors of sample size and number of items with larger sample sizes (500 or 1000) or bigger number of items (10) led to higher detection rates.

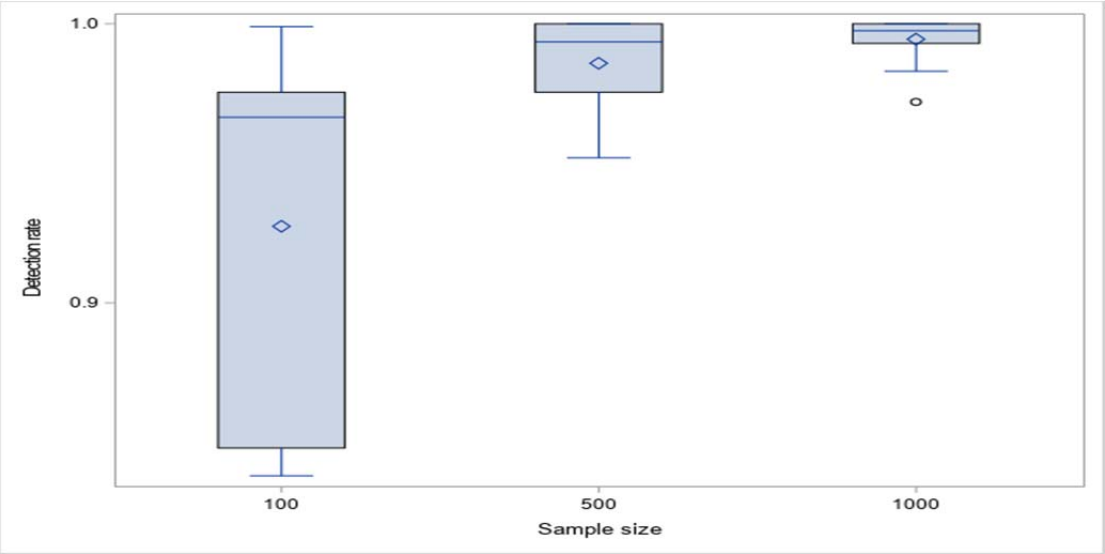


Figure 15. Distributions of detection rates of Models 1 and 2 using ΔRMSEA for scalar invariance conditions by sample size

For example while all 10-item conditions had detection rates of 95% or above, most of 5-item conditions had detection rates equal or larger than 87% (see Figures 16 and 18). In addition, as shown in Figure 19, ΔCFI also got important influence from the interaction of model and sample size with more differences of detection rates between small sample size (100) and larger sample size (500 and 1000) for Model 1 than those of Model 2. Specifically, the detection rates of both models were 100% for larger sample sizes (500 and 1000) but were smaller for Model 1 than Model 2 with small sample size (100). As seen in Table 13 and Figure 20, the detection

rates of two models were significantly different from each other only using $\Delta\chi 2$ ($\eta^2 = 0.910$) with higher rates for Model 2 than Model 1 although detection rates of Model 1 were still always equal or higher than 88%.

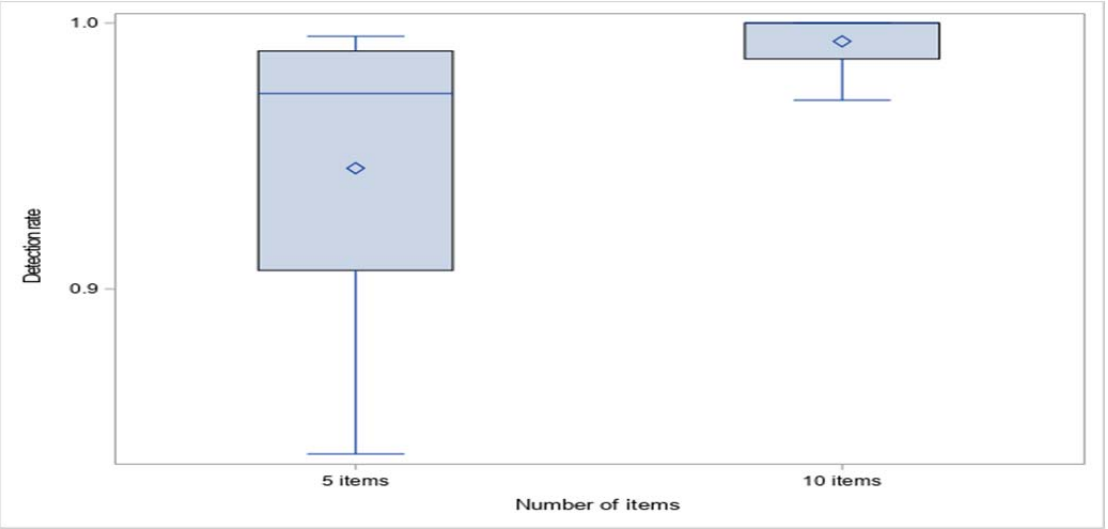


Figure 16. Distributions of detection rates of Models 1 and 2 using ΔRMSEA for scalar invariance conditions by number of items

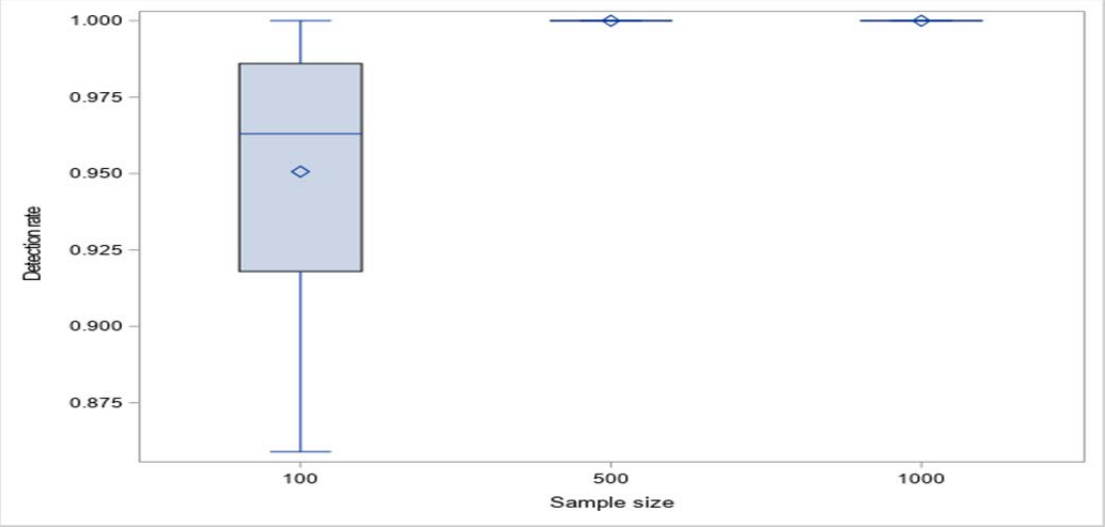


Figure 17. Distributions of detection rates of Models 1 and 2 using ΔCFI for scalar invariance conditions by sample size

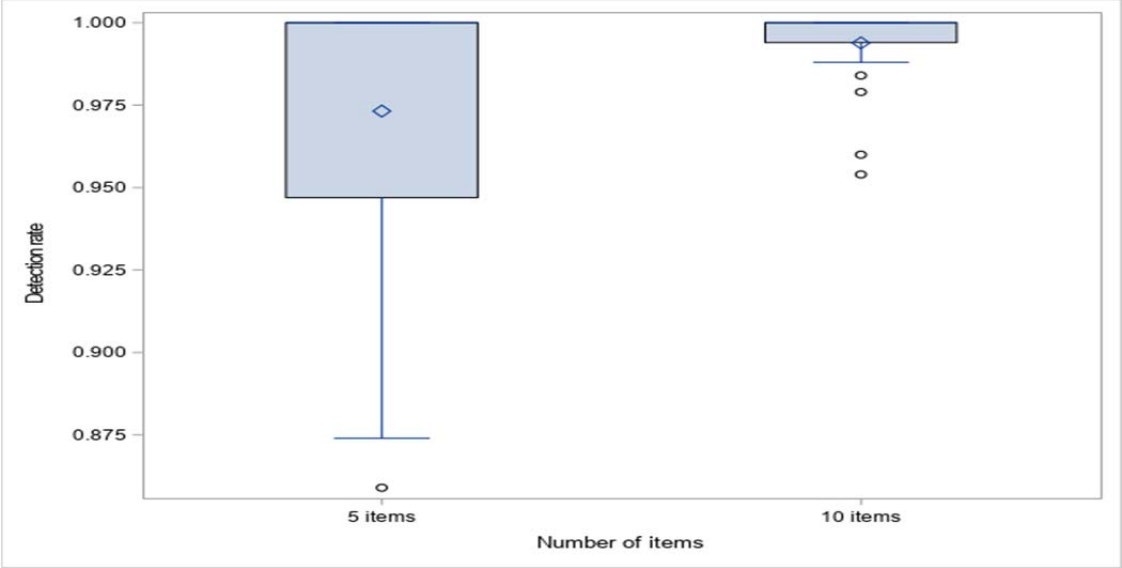


Figure 18. Distributions of detection rates of Models 1 and 2 using ΔCFI for scalar invariance conditions by number of items

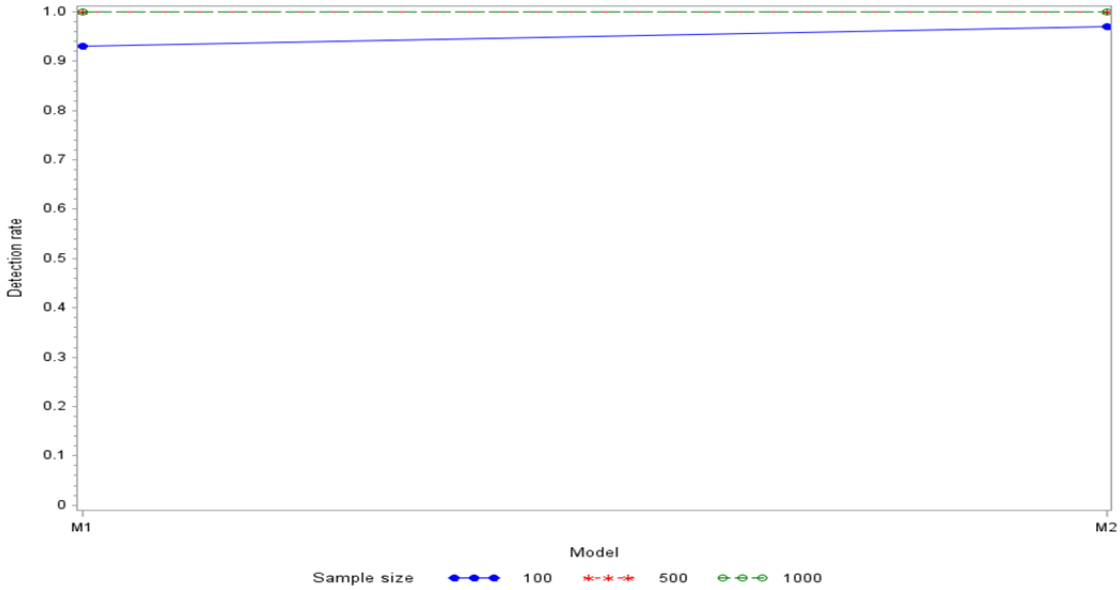


Figure 19. Distributions of detection rates of Models 1 and 2 using ΔCFI for scalar invariance conditions by model and sample size

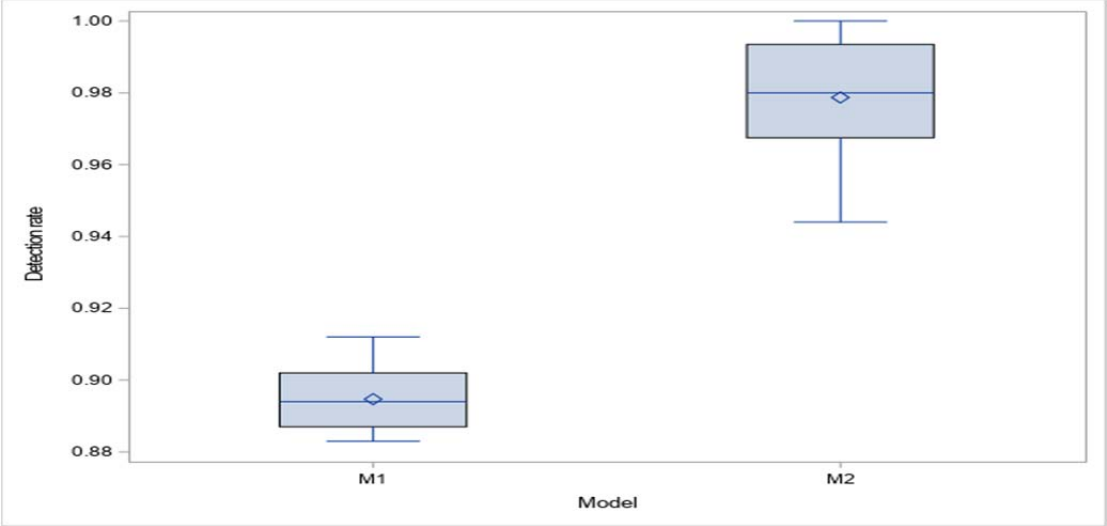


Figure 20. Distributions of detection rates of Models 1 and 2 using Δχ2 test for scalar invariance conditions by type of model

2. Result of Study 2

2.1. Detection rates of Model 3, Model 4 and Model 5

This section presents results to answer research question 1 of Study 2, i.e. how well Models 3, 4, and 5 detect measurement invariance levels across the simulation factors examined in this study measured by the detection rates for each level of MI using one of three criteria: Δχ2 or Satorra–Bentler LRT test, ΔCFI and ΔRMSEA tests. The detection rates of Model 3 and Model 5 using Δχ2 test were calculated but there were many cases of negative values of this test for the two models, especially for scalar invariance conditions. For example for the condition of five items + small factor correlation + small cluster size + small number of clusters + small ICC + small error correlation, out of 1000 replications, there were 168 and 289 replications with negative values of Δχ2 between scalar and metric invariance models for Model 3 and Model 5, respectively. The highest cases of negative chi-square difference values out of 1000 replications were 199 for Model 3 and 524 for Model 5. On the other hand, there were almost no negative values (only one condition with 11 replications out of 1000 replications) of detection rates of

Models 3 and 5 using the SB LRT test. Thus only the detection rates using SB LRT test was reported for these two models with MLR estimator.

2.1.1. Detection Rates for Configural Invariance Conditions

Table 13 shows detection rates of Models 3, 4, and 5 for configural invariance with 5-item (i.e. noninvariance in both intercepts and factor loadings between two factors and 5 items as number of items per factor) with small factor correlation. Overall the detection rates of Model 3 were often higher than those of Model 4 and Model 5 for configural invariance 5-item conditions if using $\Delta\chi 2$, Satorra–Bentler LRT or ΔCFI test. If using ΔRMSEA criterion, the detection rates for these 5-item configural invariance conditions were highest for Model 4, following by Model 5 and lowest for Model 3 with small magnitude of noninvariance but those rates were often highest for Model 3 for large noninvariance conditions.

When magnitude of noninvariance was large, all three models were able to detect configural invariance very well using $\Delta\chi 2$ or Satorra–Bentler LRT test with the detection rates were always 99% - 100% for Model 3, 84%-100% for Model 4 and 61%-100% for Model 5. The only cases where Models 4 and 5 had lower detection rates using either of the two LRT tests were large ICC combined with small cluster sizes. For ΔCFI or ΔRMSEA, the detection rates of Model 3 for large noninvariance size were higher than those rates of Model 4 and the lowest rates were often from Model 5.

When magnitude of noninvariance was small but total sample size was big (i.e. large number of clusters or small number of clusters but coupled with large cluster size), Model 3 was still able to catch the noninvariance in both factor loadings and intercepts 94% to 100% of the time if using Sattorra-Bentler LRT test. Even with small total sample size (i.e. small cluster size together with small number of clusters), Model 3 could detect configural invariance 66% to 90%

but Models 4 and 5 could only do the task 24% - 57% among 1000 replications using $\Delta\chi 2$ test or Sattorra-Bentler LRT test. The detection rates of all three models were low for small noninvariance if using ΔCFI or ΔRMSEA, except conditions of large sample sizes (i.e. cluster size of 20 and number of clusters of 80) coupled with large error correlation (i.e. 0.6) could lead to higher detection rates of 78% to 82% using ΔRMSEA.

Similar to 5-item configural invariance conditions, the detection rates of Model 3 were highest (91% - 100%), followed by Model 5 (47% - 100%) and Model 4 (69% - 100%) if using $\Delta\chi 2$ test or Sattorra-Bentler LRT test (see Table 15) across all 10-item configural invariance conditions. However the ability to detect small noninvariance in these conditions of all three models were really low if using ΔRMSEA and ΔCFI (37% or smaller). On the other hand, when the magnitude of noninvariance was large, the detection rates of three models were always 87% to 100% if using ΔCFI but were lower if using ΔRMSEA, especially for conditions with small ICC combined with small number of clusters.

2.1.2. Detection Rates for Metric Invariance Conditions

The detection rates of Models 3, 4, and 5 for metric invariance with 5-item conditions (i.e. conditions with invariant factor loadings but there was noninvariance in intercepts between two factors and the number of items per factor was five) using four criteria are presented in Table 16. When the magnitude of noninvariance was large, the ability to detect metric invariance was highest for Model 4, following by Model 5 and Model 3 using Satorra–Bentler LRT or regular LRT and the detection rates of three models were much higher for conditions with smaller ICC than those with larger ICC. However if using ΔRMSEA or ΔCFI, while detection rates of Model 4 for metric invariance with 5-item were often high (61% - 100%), these rates of Model 5 was pretty lower and the rates of Model 3 were smallest and even less than 30% for small noninvariance condtions.

Table 14. Detection rate (DR) of Models 3, 4, and 5 for configural invariance with 5-item and small factor correlation conditions

DR using Δχ2 test	DR using SB LRT		DR using ΔCFI			DR using ΔRMSEA			# of Items	NI size	Factor corr	cluster size	# of cluster	ICC	Error corr
M4	M3	M5	M3	M4	M5	M3	M4	M5							
0.45	0.61	0.45	0.33	0.26	0.23	0.13	0.41	0.31	5	Small	0.3	10	30	0.13	0.3
0.42	0.82	0.57	0.36	0.22	0.25	0.26	0.39	0.39	5	Small	0.3	10	30	0.13	0.6
0.47	0.61	0.24	0.32	0.35	0.21	0.13	0.29	0.15	5	Small	0.3	10	30	0.33	0.3
0.45	0.81	0.25	0.36	0.33	0.21	0.23	0.27	0.16	5	Small	0.3	10	30	0.33	0.6
0.88	0.98	0.90	0.28	0.19	0.20	0.27	0.67	0.47	5	Small	0.3	10	80	0.13	0.3
0.89	1.00	0.97	0.34	0.18	0.22	0.53	0.68	0.64	5	Small	0.3	10	80	0.13	0.6
0.71	0.98	0.47	0.27	0.22	0.12	0.27	0.33	0.16	5	Small	0.3	10	80	0.33	0.3
0.70	1.00	0.51	0.34	0.21	0.12	0.51	0.32	0.20	5	Small	0.3	10	80	0.33	0.6
0.77	0.94	0.74	0.34	0.23	0.24	0.18	0.55	0.37	5	Small	0.3	20	30	0.13	0.3
0.76	1.00	0.84	0.38	0.20	0.32	0.38	0.53	0.49	5	Small	0.3	20	30	0.13	0.6
0.71	0.94	0.30	0.34	0.38	0.20	0.18	0.24	0.13	5	Small	0.3	20	30	0.33	0.3
0.70	1.00	0.31	0.37	0.37	0.21	0.37	0.24	0.14	5	Small	0.3	20	30	0.33	0.6
1.00	1.00	1.00	0.27	0.15	0.21	0.46	0.81	0.66	5	Small	0.3	20	80	0.13	0.3
1.00	1.00	1.00	0.34	0.13	0.30	0.78	0.82	0.78	5	Small	0.3	20	80	0.13	0.6
0.92	1.00	0.60	0.27	0.20	0.13	0.44	0.32	0.14	5	Small	0.3	20	80	0.33	0.3
0.93	1.00	0.63	0.35	0.19	0.14	0.78	0.34	0.15	5	Small	0.3	20	80	0.33	0.6
0.96	1.00	0.96	0.97	0.89	0.86	0.73	0.93	0.85	5	Large	0.3	10	30	0.13	0.3
0.97	1.00	0.99	0.99	0.90	0.92	0.93	0.93	0.94	5	Large	0.3	10	30	0.13	0.6
0.85	0.99	0.61	0.97	0.80	0.59	0.72	0.65	0.46	5	Large	0.3	10	30	0.33	0.3
0.84	1.00	0.64	0.99	0.78	0.62	0.90	0.63	0.51	5	Large	0.3	10	30	0.33	0.6
1.00	1.00	1.00	1.00	0.99	0.99	0.98	1.00	1.00	5	Large	0.3	10	80	0.13	0.3
1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	5	Large	0.3	10	80	0.13	0.6
0.99	1.00	0.97	1.00	0.89	0.80	0.98	0.87	0.76	5	Large	0.3	10	80	0.33	0.3
0.99	1.00	0.98	1.00	0.89	0.83	1.00	0.87	0.79	5	Large	0.3	10	80	0.33	0.6
1.00	1.00	1.00	0.99	0.98	0.98	0.96	0.99	0.97	5	Large	0.3	20	30	0.13	0.3
1.00	1.00	1.00	0.99	0.98	0.99	1.00	0.99	0.99	5	Large	0.3	20	30	0.13	0.6
0.98	1.00	0.74	0.99	0.87	0.66	0.96	0.69	0.51	5	Large	0.3	20	30	0.33	0.3
0.98	1.00	0.75	0.99	0.87	0.69	1.00	0.68	0.51	5	Large	0.3	20	30	0.33	0.6
1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	5	Large	0.3	20	80	0.13	0.3
1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	5	Large	0.3	20	80	0.13	0.6
1.00	1.00	1.00	0.99	0.95	0.89	1.00	0.91	0.79	5	Large	0.3	20	80	0.33	0.3
1.00	1.00	1.00	0.99	0.95	0.90	1.00	0.92	0.80	5	Large	0.3	20	80	0.33	0.6

Note: NI size = magnitude of noninvariance, M3=Model 3, M4=Model 4, M5=Model 5, Factor Corr = factor correlation, Error Corr = error correlation

Table 15. Detection rate (DR) of Models 3, 4, and 5 for configural invariance with 10-item and small factor correlation conditions

DR using $\Delta\chi 2$ test	DR using SB LRT		DR using ΔCFI			DR using ΔRMSEA			# of items	Factor Corr	NI size	Cluster size	# of cluster	ICC	Error Corr
M4	M3	M5	M3	M4	M5	M3	M4	M5							
0.78	0.91	0.82	0.29	0.21	0.23	0.00	0.21	0.12	10	0.3	Small	10	30	0.13	0.3
0.76	0.99	0.92	0.37	0.19	0.28	0.01	0.22	0.17	10	0.3	Small	10	30	0.13	0.6
0.71	0.91	0.47	0.28	0.28	0.20	0.00	0.01	0.02	10	0.3	Small	10	30	0.33	0.3
0.69	0.98	0.53	0.37	0.25	0.20	0.00	0.01	0.02	10	0.3	Small	10	30	0.33	0.6
1.00	1.00	1.00	0.19	0.11	0.13	0.03	0.40	0.21	10	0.3	Small	10	80	0.13	0.3
1.00	1.00	1.00	0.29	0.09	0.17	0.11	0.40	0.38	10	0.3	Small	10	80	0.13	0.6
0.98	1.00	0.93	0.18	0.10	0.07	0.02	0.01	0.01	10	0.3	Small	10	80	0.33	0.3
0.97	1.00	0.96	0.29	0.09	0.07	0.11	0.01	0.02	10	0.3	Small	10	80	0.33	0.6
0.98	1.00	0.98	0.24	0.16	0.21	0.00	0.15	0.14	10	0.3	Small	20	30	0.13	0.3
0.98	1.00	1.00	0.34	0.13	0.29	0.01	0.14	0.20	10	0.3	Small	20	30	0.13	0.6
0.94	1.00	0.67	0.24	0.24	0.17	0.00	0.00	0.01	10	0.3	Small	20	30	0.33	0.3
0.93	1.00	0.71	0.34	0.21	0.18	0.01	0.00	0.01	10	0.3	Small	20	30	0.33	0.6
1.00	1.00	1.00	0.14	0.06	0.11	0.03	0.36	0.24	10	0.3	Small	20	80	0.13	0.3
1.00	1.00	1.00	0.24	0.05	0.18	0.19	0.33	0.42	10	0.3	Small	20	80	0.13	0.6
1.00	1.00	0.99	0.13	0.06	0.05	0.02	0.00	0.00	10	0.3	Small	20	80	0.33	0.3
1.00	1.00	1.00	0.24	0.05	0.05	0.18	0.00	0.00	10	0.3	Small	20	80	0.33	0.6
1.00	1.00	1.00	1.00	0.98	0.96	0.19	0.95	0.82	10	0.3	Large	10	30	0.13	0.3
1.00	1.00	1.00	1.00	0.98	0.99	0.57	0.94	0.93	10	0.3	Large	10	30	0.13	0.6
1.00	1.00	0.96	1.00	0.94	0.87	0.12	0.29	0.25	10	0.3	Large	10	30	0.33	0.3
1.00	1.00	0.98	1.00	0.94	0.90	0.48	0.28	0.25	10	0.3	Large	10	30	0.33	0.6
1.00	1.00	1.00	1.00	1.00	1.00	0.83	1.00	1.00	10	0.3	Large	10	80	0.13	0.3
1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	10	0.3	Large	10	80	0.13	0.6
1.00	1.00	1.00	1.00	1.00	0.97	0.79	0.67	0.60	10	0.3	Large	10	80	0.33	0.3
1.00	1.00	1.00	1.00	1.00	0.98	0.99	0.67	0.68	10	0.3	Large	10	80	0.33	0.6
1.00	1.00	1.00	0.89	1.00	1.00	0.57	0.98	0.93	10	0.3	Large	20	30	0.13	0.3
1.00	1.00	1.00	0.89	1.00	1.00	0.96	0.98	0.98	10	0.3	Large	20	30	0.13	0.6
1.00	1.00	1.00	0.89	0.98	0.93	0.50	0.22	0.23	10	0.3	Large	20	30	0.33	0.3
1.00	1.00	1.00	0.89	0.99	0.95	0.94	0.19	0.21	10	0.3	Large	20	30	0.33	0.6
1.00	1.00	1.00	0.87	1.00	1.00	0.99	1.00	1.00	10	0.3	Large	20	80	0.13	0.3
1.00	1.00	1.00	0.87	1.00	1.00	1.00	1.00	1.00	10	0.3	Large	20	80	0.13	0.6
1.00	1.00	1.00	0.87	1.00	0.99	0.99	0.69	0.54	10	0.3	Large	20	80	0.33	0.3
1.00	1.00	1.00	0.87	1.00	0.99	1.00	0.68	0.54	10	0.3	Large	20	80	0.33	0.6

Note: NI size = magnitude of noninvariance, M3=Model 3, M4=Model 4, M5=Model 5, Factor Corr = factor correlation, Error Corr = error correlation

Increasing number of items per factor to ten results in higher detection rates of metric invariance using $\Delta\chi 2$ test or Satorra–Bentler LRT for all three models even with small magnitude of noninvariance (see Table 17). While the detection rates of Model 3 and Model 5 were always higher than 90% across all metric invariance conditions, these rates of Model 4 were only above 90% for small ICC conditions and ranged from 63% to 86% for large ICC condition. The detection rates using ΔRMSEA or ΔCFI also improved when the number of items per factor went up to ten. While the ability to detect metric invariance with 10-item was really high to perfect for Models 4 and 5 using these two criteria, that ability of Model 3 was only high for a half of conditions (small ICC + small cluster size + small noninvariance or small cluster size + large noninvariance or small ICC + large noninvariance) if using ΔCFI and even fewer conditions (e.g. small cluster size + large number of clusters + small ICC) if using ΔRMSEA. Model 3 could hardly detect 10-item metric invariance for conditions of small number of clusters + large ICC if using ΔRMSEA or conditions of large cluster size + large ICC if using ΔCFI.

2.1.3. Detection Rates for Scalar Invariance Conditions

The detection rates of Models 3, 4, and 5 for scalar invariance (i.e. invariant intercepts and factor loadings) with small factor correlations conditions using chi-square difference test and suggested difference cutoffs of CFI and RMSEA difference tests are presented in Table 18. While the detection rates for scalar invariance conditions using Satorra-Bentler LRT test for Models 3 and 5 were always high and pretty similar to each other and across conditions (84%-89% for Model 5 and 87% - 91% for Model 3), those rates of Model 4 were much lower (0.01% - 74%), particularly for higher degree of ICC conditions with the rates below 28%.

Table 16. Detection rate of Models 3, 4, 5 for metric invariance with 5 items and small factor correlation

DR using $\Delta\chi 2$ test	Detection rate using SB LRT		Detection rate using ΔCFI			Detection rate using ΔRMSEA			# of Items	Factor Corr	NI size	Cluster size	# of clusters	ICC	Error Corr
M4	M3	M5	M3	M4	M5	M3	M4	M5							
0.89	0.63	0.73	0.30	0.86	0.77	0.13	0.85	0.75	5	0.3	Small	10	30	0.13	0.3
0.91	0.66	0.75	0.20	0.86	0.82	0.14	0.86	0.80	5	0.3	Small	10	30	0.13	0.6
0.69	0.26	0.32	0.06	0.73	0.47	0.03	0.61	0.39	5	0.3	Small	10	30	0.33	0.3
0.69	0.26	0.31	0.02	0.72	0.48	0.03	0.63	0.42	5	0.3	Small	10	30	0.33	0.6
0.94	0.93	0.93	0.22	0.98	0.96	0.29	0.96	0.97	5	0.3	Small	10	80	0.13	0.3
0.96	0.94	0.93	0.08	0.98	0.98	0.32	0.96	0.98	5	0.3	Small	10	80	0.13	0.6
0.75	0.59	0.67	0.01	0.87	0.68	0.03	0.80	0.64	5	0.3	Small	10	80	0.33	0.3
0.77	0.62	0.67	0.00	0.87	0.69	0.03	0.81	0.65	5	0.3	Small	10	80	0.33	0.6
0.91	0.69	0.79	0.03	0.93	0.83	0.05	0.90	0.81	5	0.3	Small	20	30	0.13	0.3
0.93	0.71	0.79	0.01	0.93	0.86	0.07	0.92	0.86	5	0.3	Small	20	30	0.13	0.6
0.54	0.25	0.33	0.01	0.77	0.49	0.01	0.66	0.39	5	0.3	Small	20	30	0.33	0.3
0.55	0.26	0.31	0.00	0.78	0.49	0.01	0.67	0.38	5	0.3	Small	20	30	0.33	0.6
0.93	0.95	0.95	0.00	0.98	0.97	0.09	0.99	0.98	5	0.3	Small	20	80	0.13	0.3
0.94	0.96	0.94	0.00	0.98	0.98	0.09	0.99	0.99	5	0.3	Small	20	80	0.13	0.6
0.59	0.64	0.69	0.00	0.91	0.68	0.00	0.85	0.59	5	0.3	Small	20	80	0.33	0.3
0.60	0.64	0.68	0.00	0.91	0.69	0.00	0.84	0.59	5	0.3	Small	20	80	0.33	0.6
0.94	0.94	0.93	0.96	0.99	0.99	0.77	0.92	0.96	5	0.3	Large	10	30	0.13	0.3
0.96	0.93	0.92	0.95	0.99	0.99	0.79	0.94	0.96	5	0.3	Large	10	30	0.13	0.6
0.77	0.76	0.81	0.50	0.86	0.88	0.21	0.85	0.85	5	0.3	Large	10	30	0.33	0.3
0.78	0.76	0.79	0.34	0.86	0.88	0.21	0.87	0.85	5	0.3	Large	10	30	0.33	0.6
0.94	0.95	0.94	1.00	1.00	1.00	0.99	0.96	1.00	5	0.3	Large	10	80	0.13	0.3
0.96	0.95	0.94	1.00	1.00	1.00	0.99	0.97	0.99	5	0.3	Large	10	80	0.13	0.6
0.76	0.95	0.93	0.51	0.99	1.00	0.48	0.94	0.98	5	0.3	Large	10	80	0.33	0.3
0.78	0.95	0.93	0.25	0.99	1.00	0.48	0.95	0.98	5	0.3	Large	10	80	0.33	0.6
0.91	0.94	0.93	0.81	1.00	1.00	0.48	0.94	0.99	5	0.3	Large	20	30	0.13	0.3
0.94	0.93	0.93	0.50	1.00	1.00	0.50	0.95	0.99	5	0.3	Large	20	30	0.13	0.6
0.56	0.79	0.83	0.09	0.88	0.92	0.09	0.90	0.88	5	0.3	Large	20	30	0.33	0.3
0.57	0.79	0.81	0.01	0.89	0.92	0.10	0.91	0.86	5	0.3	Large	20	30	0.33	0.6
0.93	0.96	0.95	0.90	1.00	1.00	0.90	0.99	1.00	5	0.3	Large	20	80	0.13	0.3
0.94	0.96	0.94	0.50	1.00	1.00	0.91	0.99	1.00	5	0.3	Large	20	80	0.13	0.6
0.59	0.96	0.93	0.01	1.00	1.00	0.19	0.97	0.99	5	0.3	Large	20	80	0.33	0.3
0.60	0.96	0.92	0.00	1.00	1.00	0.19	0.96	0.99	5	0.3	Large	20	80	0.33	0.6

Note: NI size = magnitude of noninvariance, M3=Model 3, M4=Model 4, M5=Model 5, Factor Corr = factor correlation, Error Corr = error correlation

Table 17. Detection rate of Models 3, 4, and 5 (M3, M4, and M5) for metric invariance with 10-item and small factor correlation

Detection rate using Δχ2 test	Detection rate using SB LRT		Detection rate using ΔCFI			Detection rate using ΔRMSEA			# of Items	Factor Corr	NI size	Cluster size	# of clusters	ICC	Error corr
M4	M3	M5	M3	M4	M5	M3	M4	M5							
0.95	0.92	0.93	1.00	1.00	1.00	0.16	1.00	1.00	10	0.3	Small	10	30	0.13	0.3
0.98	0.93	0.93	1.00	1.00	1.00	0.19	1.00	1.00	10	0.3	Small	10	30	0.13	0.6
0.80	0.92	0.92	0.83	0.98	0.99	0.01	1.00	0.97	10	0.3	Small	10	30	0.33	0.3
0.84	0.93	0.92	0.58	0.99	1.00	0.01	1.00	0.98	10	0.3	Small	10	30	0.33	0.6
0.96	0.94	0.95	1.00	1.00	1.00	0.89	1.00	1.00	10	0.3	Small	10	80	0.13	0.3
0.98	0.94	0.94	1.00	1.00	1.00	0.93	1.00	1.00	10	0.3	Small	10	80	0.13	0.6
0.83	0.94	0.95	0.95	1.00	1.00	0.19	1.00	1.00	10	0.3	Small	10	80	0.33	0.3
0.86	0.94	0.95	0.68	1.00	1.00	0.21	1.00	1.00	10	0.3	Small	10	80	0.33	0.6
0.94	0.93	0.93	0.85	1.00	1.00	0.02	1.00	1.00	10	0.3	Small	20	30	0.13	0.3
0.97	0.92	0.92	0.55	1.00	1.00	0.02	1.00	1.00	10	0.3	Small	20	30	0.13	0.6
0.65	0.93	0.92	0.05	1.00	1.00	0.00	1.00	0.95	10	0.3	Small	20	30	0.33	0.3
0.67	0.92	0.92	0.00	1.00	1.00	0.00	1.00	0.95	10	0.3	Small	20	30	0.33	0.6
0.95	0.95	0.96	0.87	1.00	1.00	0.33	1.00	1.00	10	0.3	Small	20	80	0.13	0.3
0.97	0.95	0.95	0.64	1.00	1.00	0.37	1.00	1.00	10	0.3	Small	20	80	0.13	0.6
0.63	0.94	0.95	0.01	1.00	1.00	0.01	1.00	1.00	10	0.3	Small	20	80	0.33	0.3
0.66	0.95	0.95	0.00	1.00	1.00	0.02	1.00	1.00	10	0.3	Small	20	80	0.33	0.6
0.95	0.92	0.93	1.00	1.00	1.00	0.76	1.00	1.00	10	0.3	Large	10	30	0.13	0.3
0.98	0.93	0.93	1.00	1.00	1.00	0.85	1.00	1.00	10	0.3	Large	10	30	0.13	0.6
0.80	0.92	0.92	0.99	0.98	0.99	0.03	1.00	1.00	10	0.3	Large	10	30	0.33	0.3
0.84	0.93	0.92	0.94	0.99	1.00	0.04	1.00	1.00	10	0.3	Large	10	30	0.33	0.6
0.96	0.94	0.95	1.00	1.00	1.00	1.00	1.00	1.00	10	0.3	Large	10	80	0.13	0.3
0.98	0.94	0.94	1.00	1.00	1.00	1.00	1.00	1.00	10	0.3	Large	10	80	0.13	0.6
0.83	0.94	0.95	1.00	1.00	1.00	0.54	1.00	1.00	10	0.3	Large	10	80	0.33	0.3
0.86	0.94	0.95	0.99	1.00	1.00	0.58	1.00	1.00	10	0.3	Large	10	80	0.33	0.6
0.94	0.93	0.93	0.89	1.00	1.00	0.24	1.00	1.00	10	0.3	Large	20	30	0.13	0.3
0.97	0.92	0.92	0.89	1.00	1.00	0.26	1.00	1.00	10	0.3	Large	20	30	0.13	0.6
0.65	0.93	0.92	0.39	1.00	1.00	0.00	1.00	0.99	10	0.3	Large	20	30	0.33	0.3
0.67	0.92	0.92	0.09	1.00	1.00	0.00	1.00	1.00	10	0.3	Large	20	30	0.33	0.6
0.95	0.95	0.96	0.87	1.00	1.00	0.95	1.00	1.00	10	0.3	Large	20	80	0.13	0.3
0.97	0.95	0.95	0.87	1.00	1.00	0.97	1.00	1.00	10	0.3	Large	20	80	0.13	0.6
0.63	0.94	0.95	0.40	1.00	1.00	0.12	1.00	1.00	10	0.3	Large	20	80	0.33	0.3
0.66	0.95	0.95	0.02	1.00	1.00	0.12	1.00	1.00	10	0.3	Large	20	80	0.33	0.6

Note: NI size = magnitude of noninvariance, M3=Model 3, M4=Model 4, M5=Model 5, Factor Corr = factor correlation, Error Corr = error correlation

Using ΔCFI or ΔRMSEA criteria, while Model 3 was able to detect scalar invariance all the time (99% - 100%), Model 5 could detect this level of measurement invariance 79% - 100% across all conditions, and Models 4 could perform this task 81% - 100% for only the conditions with low ICC + large number of clusters. The detection rates of Model 4 for high ICC + small number of clusters conditions were only 34% - 47% if using ΔRMSEA and 53% - 58% for high ICC + small number of clusters + small number of items conditions if using ΔCFI.

2.2. Impact of simulation factors on the detection rates for Models 3, 4 and 5

Study 2 includes eight simulation factors: 1) number of items (5 and 10); 2) magnitude of noninvariance (zero, small, large); 3) location of noninvariance (noninvariance in both intercepts and factor loadings, noninvariance in intercepts only); 4) factor correlation (0.3 and 0.5); 5) cluster size (10 and 20); 6) number of clusters (30 and 80); 7) ICC (0.13 and 0.33); and 8) error correlation (0.3 and 0.8). Effect size of each simulation factor and the type of model (i.e. Model 3, 4 or 5) on one of the three outcomes for each model (i.e. Satorra–Bentler LRT, ΔCFI, and ΔRMSEA for Models 3 and 5, and $\Delta\chi 2$ test, ΔCFI, ΔRMSEA for Model 4) were calculated using eta-squared analyses for main effects and first-degree interactions of the type of model with each simulation factor with suggested cut-off value of 0.058 as significant effect by Cohen (1992). As explained earlier, the detection rates of Models 3 and 5 using $\Delta\chi 2$ test were calculated but had negative values for several cases and were not reported in this dissertation. Instead, the detection rates of Models 3 and 5 using SB LRT and the detection rates of Model 4 using $\Delta\chi 2$ test were combined into one outcome criterion named as ΔLRT (loglikelihood ratio difference) to calculate effect sizes.

Table 18. Detection rates (DR) of Models 3, 4 and 5 (M3, M4, and M5) for scalar invariance with small factor correlation conditions

DR using $\Delta\chi2$ test	DR using Satorra-Bentler $\chi2$ test		DR using ΔCFI			DR using ΔRMSEA			# of items	Factor correlation	Cluster size	# of clusters	ICC	Error correlation
M4	M3	M5	M3	M4	M5	M3	M4	M5						
0.66	0.89	0.87	0.99	0.87	0.93	0.99	0.67	0.88	5	0.3	10	30	0.13	0.3
0.74	0.88	0.86	1.00	0.91	0.92	0.99	0.74	0.85	5	0.3	10	30	0.13	0.6
0.26	0.89	0.85	0.99	0.38	0.79	0.99	0.53	0.83	5	0.3	10	30	0.33	0.3
0.28	0.88	0.84	1.00	0.41	0.79	1.00	0.56	0.84	5	0.3	10	30	0.33	0.6
0.67	0.90	0.89	1.00	1.00	1.00	1.00	0.82	0.95	5	0.3	10	80	0.13	0.3
0.74	0.91	0.89	1.00	1.00	1.00	1.00	0.86	0.95	5	0.3	10	80	0.13	0.6
0.28	0.91	0.88	1.00	0.81	0.95	1.00	0.73	0.94	5	0.3	10	80	0.33	0.3
0.30	0.91	0.88	1.00	0.84	0.95	1.00	0.75	0.92	5	0.3	10	80	0.33	0.6
0.39	0.88	0.88	1.00	0.86	0.95	1.00	0.62	0.90	5	0.3	20	30	0.13	0.3
0.44	0.88	0.87	1.00	0.88	0.94	1.00	0.66	0.89	5	0.3	20	30	0.13	0.6
0.08	0.88	0.85	1.00	0.34	0.83	1.00	0.56	0.88	5	0.3	20	30	0.33	0.3
0.09	0.88	0.85	1.00	0.34	0.81	1.00	0.58	0.89	5	0.3	20	30	0.33	0.6
0.42	0.91	0.89	1.00	0.99	1.00	1.00	0.81	0.97	5	0.3	20	80	0.13	0.3
0.47	0.91	0.89	1.00	0.99	0.99	1.00	0.82	0.96	5	0.3	20	80	0.13	0.6
0.09	0.90	0.87	1.00	0.80	0.96	1.00	0.78	0.96	5	0.3	20	80	0.33	0.3
0.09	0.91	0.87	1.00	0.81	0.96	1.00	0.78	0.96	5	0.3	20	80	0.33	0.6
0.49	0.87	0.88	1.00	0.94	0.98	1.00	0.92	0.98	10	0.3	10	30	0.13	0.3
0.63	0.87	0.87	1.00	0.98	0.98	1.00	0.95	0.98	10	0.3	10	30	0.13	0.6
0.10	0.86	0.86	1.00	0.42	0.89	1.00	0.94	0.98	10	0.3	10	30	0.33	0.3
0.13	0.87	0.86	1.00	0.47	0.90	1.00	0.95	0.98	10	0.3	10	30	0.33	0.6
0.47	0.88	0.88	1.00	1.00	1.00	1.00	0.99	1.00	10	0.3	10	80	0.13	0.3
0.59	0.88	0.88	1.00	1.00	1.00	1.00	1.00	1.00	10	0.3	10	80	0.13	0.6
0.10	0.89	0.88	1.00	0.91	0.99	1.00	1.00	1.00	10	0.3	10	80	0.33	0.3
0.11	0.88	0.89	1.00	0.93	0.99	1.00	1.00	1.00	10	0.3	10	80	0.33	0.6
0.16	0.87	0.87	1.00	0.93	0.99	1.00	0.94	0.99	10	0.3	20	30	0.13	0.3
0.22	0.87	0.86	1.00	0.96	0.99	1.00	0.95	0.98	10	0.3	20	30	0.13	0.6
0.01	0.87	0.86	1.00	0.37	0.91	1.00	0.95	0.99	10	0.3	20	30	0.33	0.3
0.01	0.87	0.86	1.00	0.40	0.90	1.00	0.96	0.99	10	0.3	20	30	0.33	0.6
0.17	0.90	0.89	1.00	1.00	1.00	1.00	0.99	1.00	10	0.3	20	80	0.13	0.3
0.21	0.89	0.89	1.00	1.00	1.00	1.00	1.00	1.00	10	0.3	20	80	0.13	0.6
0.01	0.89	0.89	1.00	0.92	0.99	1.00	1.00	1.00	10	0.3	20	80	0.33	0.3
0.01	0.89	0.89	1.00	0.93	0.99	1.00	1.00	1.00	10	0.3	20	80	0.33	0.6

2.2.1. Effect Sizes for Configural Invariance Conditions

Table 19 presented eta-squared values of significant simulation factors for configural invariance using each of the three outcome criteria.

Table 19. Effect sizes of significant factors on detection rates of Models 3, 4 and 5 for configural invariance

Criteria	Model	# of items	# of clusters	Magnitude of noninvariance	ICC
ΔRMSEA		0.09	0.06	0.58	0.11
ΔCFI				0.94	
ΔLRT	0.09	0.09	0.11	0.17	

As seen in Table 19, the simulation factor of magnitude of noninvariance had significant effect on all of three outcome criteria with strongest effect on ΔCFI (η^2= 0.94) and least strong on ΔLRT (η^2=0.17). Figures 23, 25, and 29 show that bigger degree of noninvariance resulted in much higher detection rates than smaller noninvariance although the differences were larger using ΔCFI than if using ΔRMSEA or either of the two LRT tests (i.e. Satorra–Bentler LRT for Models 3 and 5 and regular $\Delta\chi 2$ test for Model 4). Both the number of items and number of clusters factors significantly impacted on the detection rates for configural invariance using ΔRMSEA or ΔLRT tests with larger number of items or number of clusters led to higher detection rates (see Figures 21, 22 and Figures 27, 28).

ICC played important role on detection rates for configural invariance of three models in Study 2 only when using ΔRMSEA criterion. The smaller the ICC was, the higher the detection rates for configural invariance using this criterion (see Figure 24). The detection rates of three models for configural invariance were significantly different only when using ΔLRT (Figure 26) and were pretty similar if using other two other alternative fit criteria.

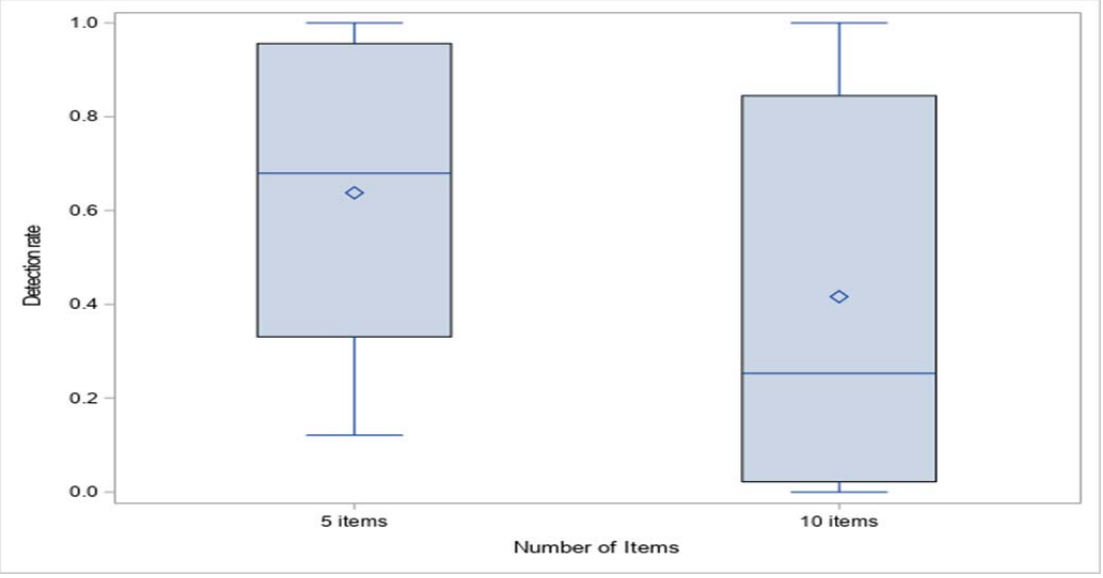


Figure 21. Distributions of detection rates of three models with partially nested data for configural invariance using ΔRMSEA by number of items

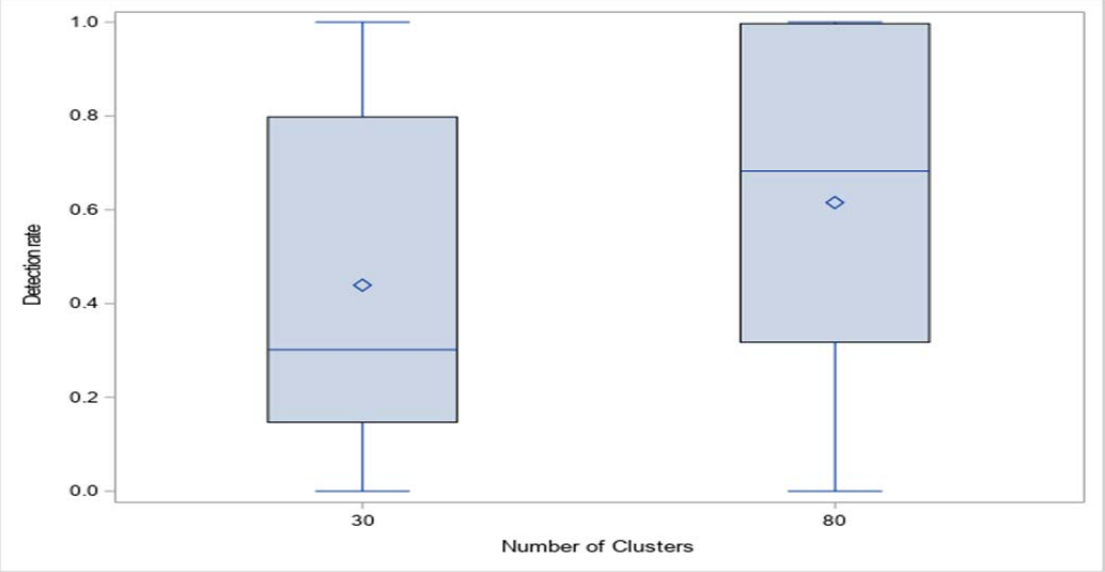


Figure 22. Distributions of detection rates of three models with partially nested data for configural invariance using ΔRMSEA by number of clusters

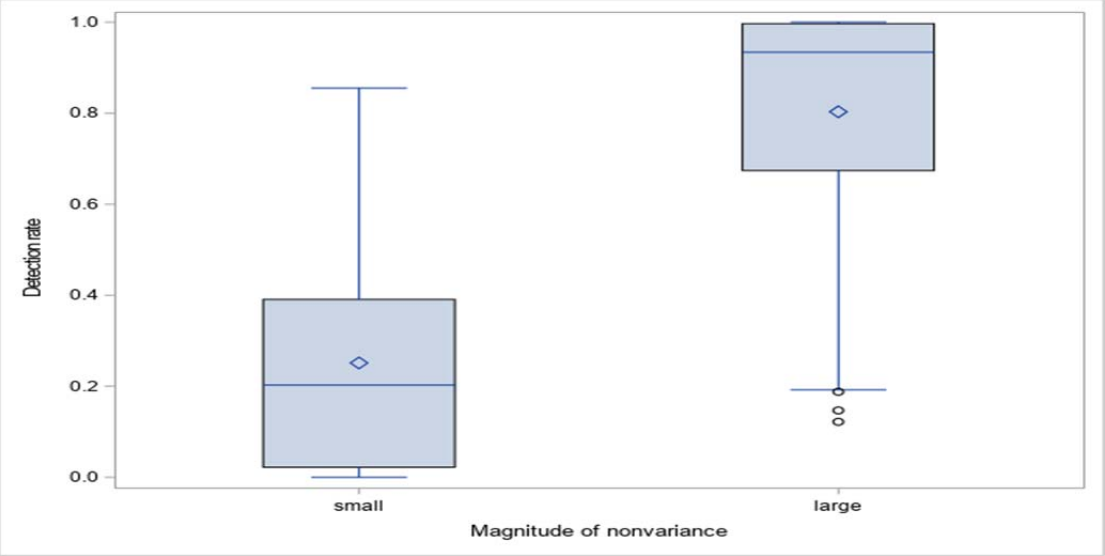


Figure 23. Distributions of detection rates of three models with partially nested data for configural invariance using ΔRMSEA by magnitude of noninvariance

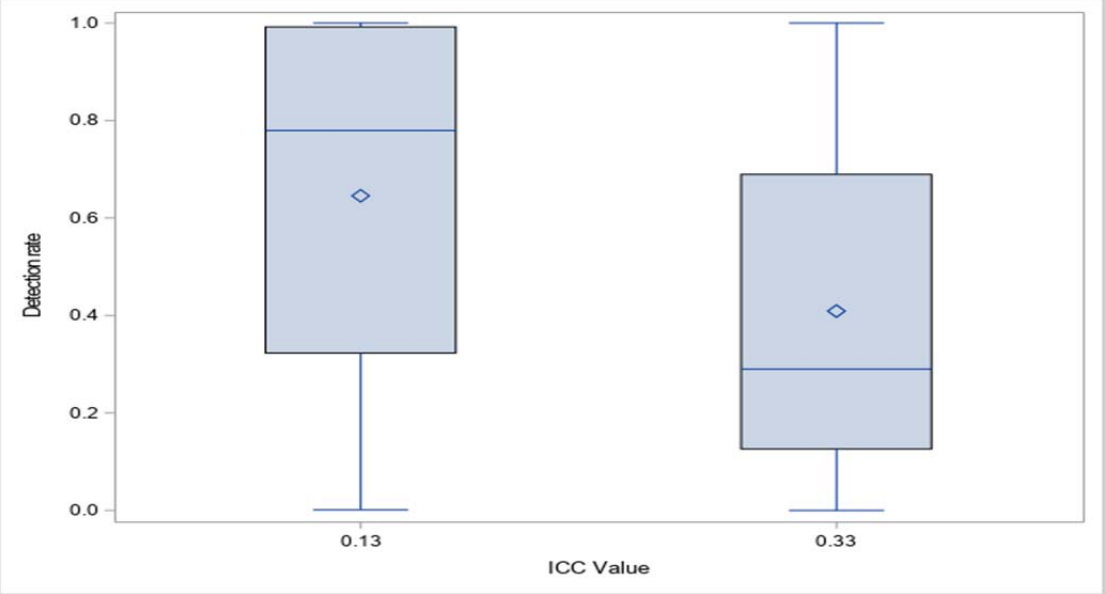


Figure 24. Distributions of detection rates of three models with partially nested data for configural invariance using ΔRMSEA by ICC

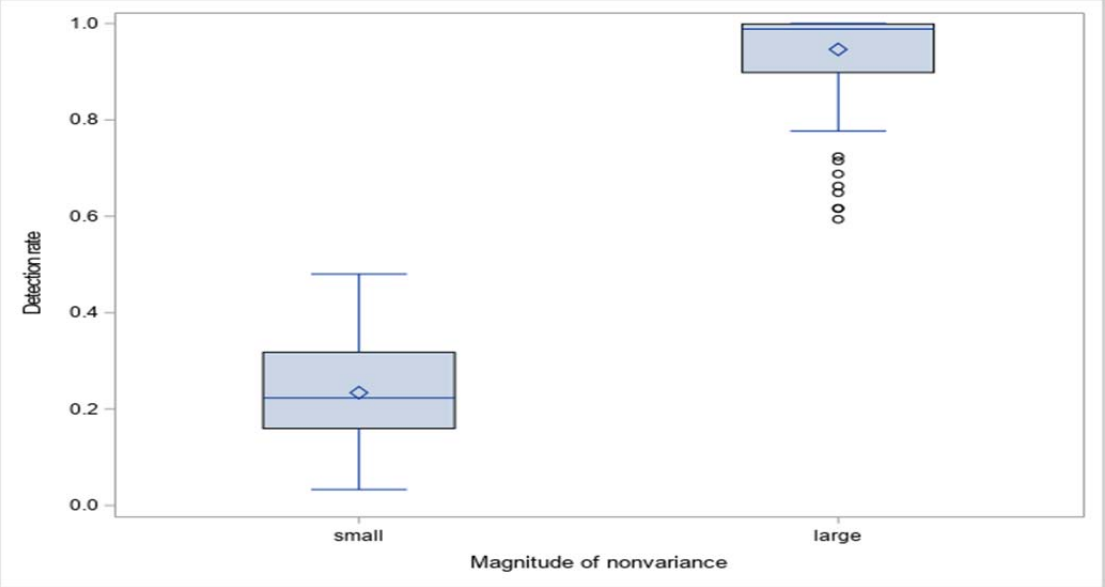


Figure 25. Distributions of detection rates of three models with partially nested data for configural invariance using ΔCFI by magnitude of noninvariance

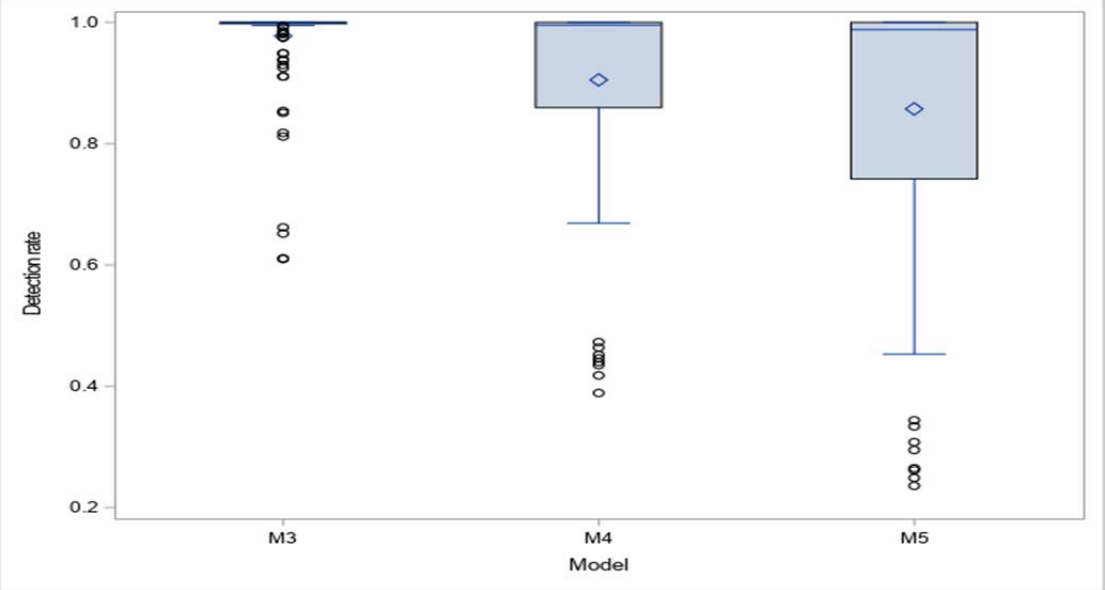


Figure 26. Distributions of detection rates of three models with partially nested data for configural invariance using $\Delta\chi 2$ or SB LRT test by model

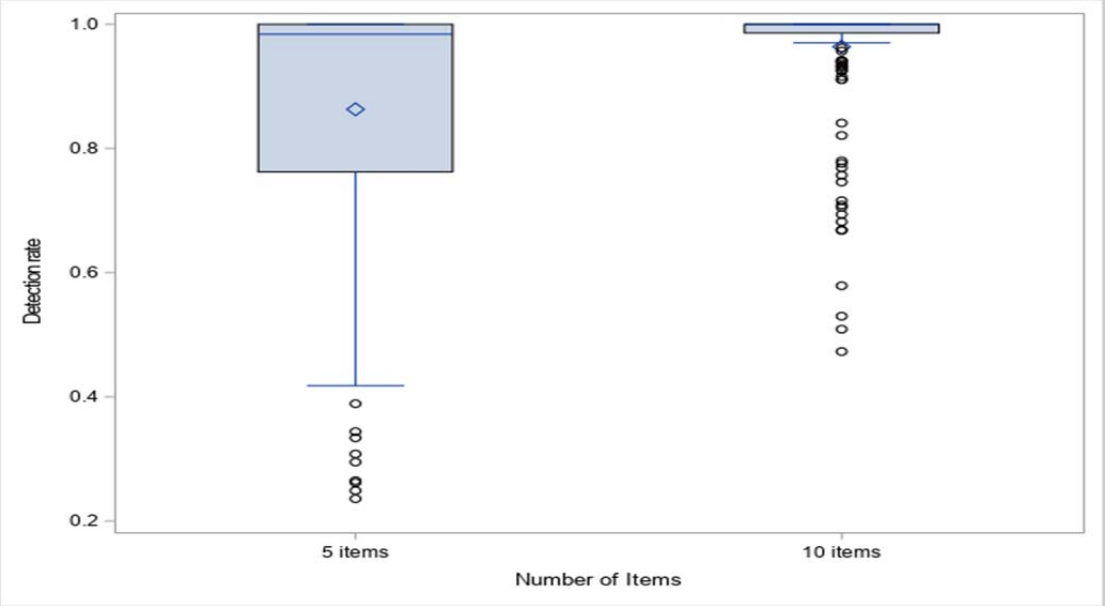


Figure 27. Distributions of detection rates of three models with partially nested data for configural invariance using Δχ2 or SB LRT test by number of items

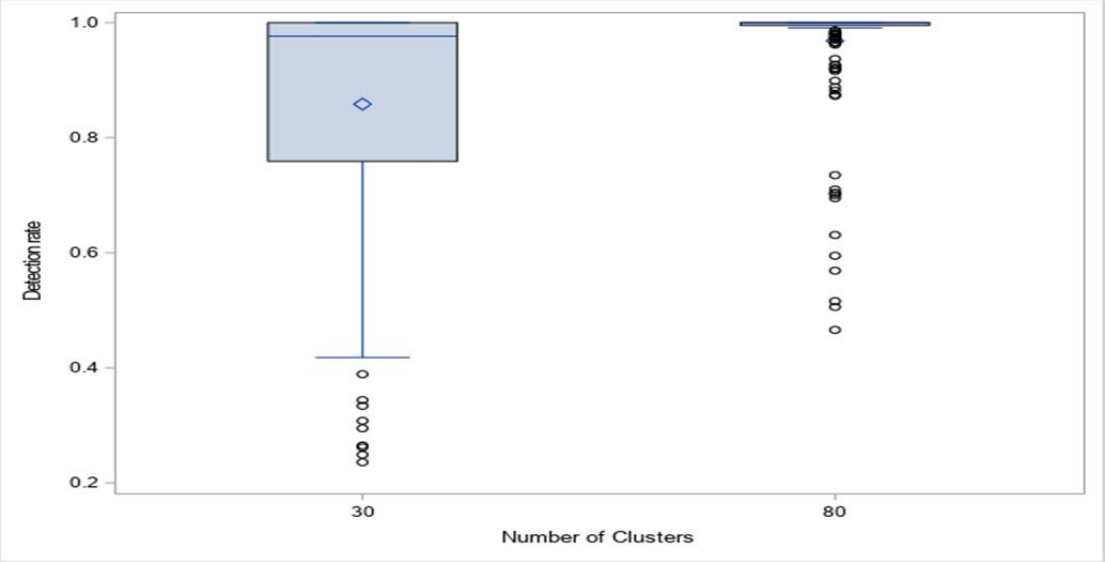


Figure 28. Distributions of detection rates of three models with partially nested data for configural invariance using Δχ2 or SB LRT test by number of clusters

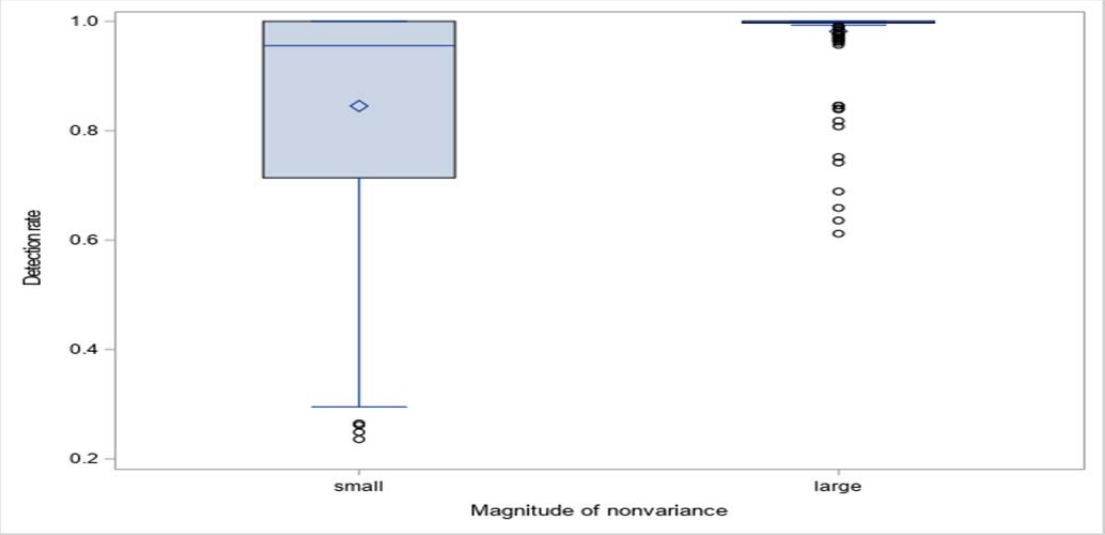


Figure 29. Distributions of detection rates of three models with partially nested data for configural invariance using Δχ2 or SB LRT test by magnitude of noninvariance

2.2.2. Effect sizes for Metric Invariance Conditions

Effect sizes of the factors that played important role on the detection rates using three outcome criteria for metric invariance conditions are shown in Table 20. The ICC significantly impacted on the detection rates of metric invariance conditions for all three outcome criteria with larger ICC resulted in lower detection rates (see Figures 32, 36 and 39).

Table 20: *Effect sizes of significant factors on detection rates of Models 3, 4, and 5 for metric invariance*

	Model	# of items	Magnitude of noninvariance	Model * Cluster size	ICC
ΔRMSEA	0.61		0.06		0.07
ΔCFI	0.41	0.10		0.06	0.07
ΔLRT		0.15	0.07		0.22

As seen in Figures 30 and 33, while the detection rates of metric invariance using ΔRMSEA or ΔCFI were similar for Models 4 and 5, these rates were lower with larger variance for Model 3 than Models 4 and 5. However, it should be kept in mind that the cutoff of RMSEA difference (and CFI difference in the following section) was developed on the basis of a single level model so it is relevant to Models 4 and 5 but may not be appropriate to Model 3 which is multilevel.

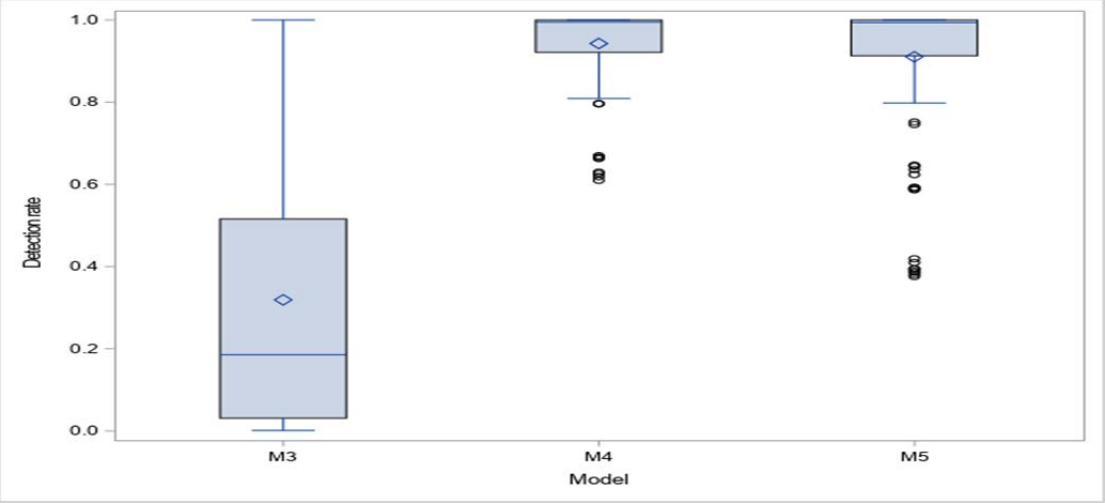


Figure 30. Distributions of detection rates of three models with partially nested data for metric invariance using ΔRMSEA by model

While number of items had significant effects on detection rates using ΔCFI and ΔLRT, magnitude of noninvariance was the significant factor on the detection rates using ΔRMSEA and ΔLRT. As shown in Figures 31 and 38, while majority of metric invariance conditions had detection rates of 80% or higher with large noninvariance using both ΔRMSEA and ΔLRT, about only a half of metric invariance conditions had detection rates of 60% or above if using ΔRMSEA and 80% or above if using ΔLRT with small noninvariance.

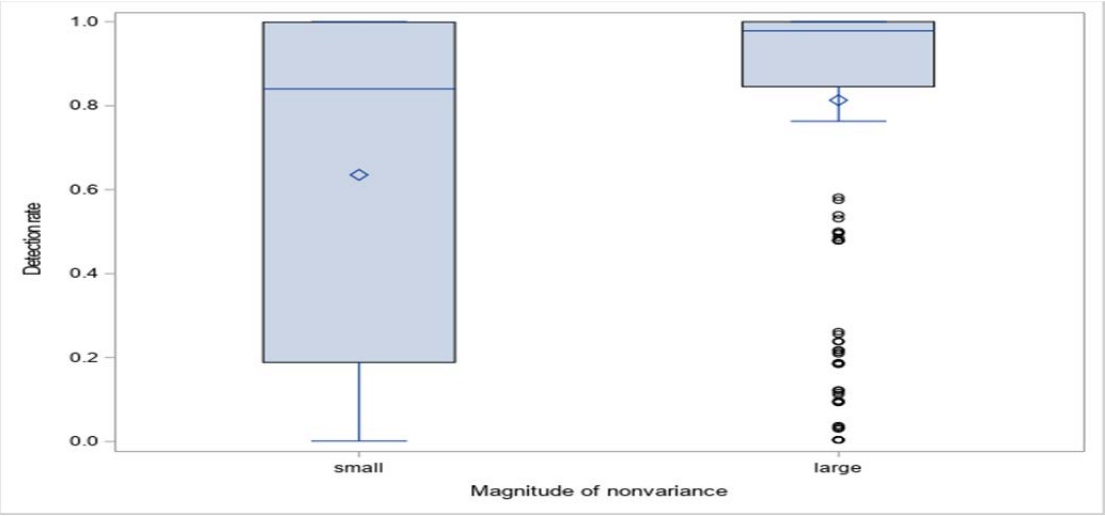


Figure 31. Distributions of detection rates of three models with partially nested data for metric invariance using ΔRMSEA by magnitude of noninvariance

As presented in Figures 34 and 37, while all of metric invariance conditions had detection rates of 60% or above (with mean =90%) using ΔLRT and many invariance conditions had detection rates 80% or higher with average rates of 90% using ΔCFI for 10-item, the average of detection rates for 5-items were only nearly 70% for ΔCFI and 80% for ΔLRT.

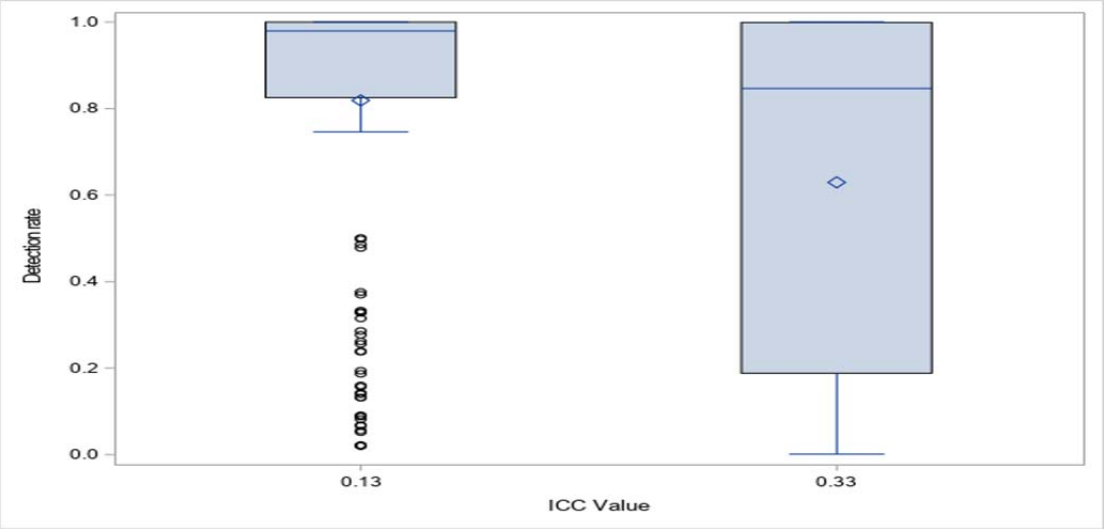


Figure 32. Distributions of detection rates of three models with partially nested data for metric invariance using ΔRMSEA by ICC

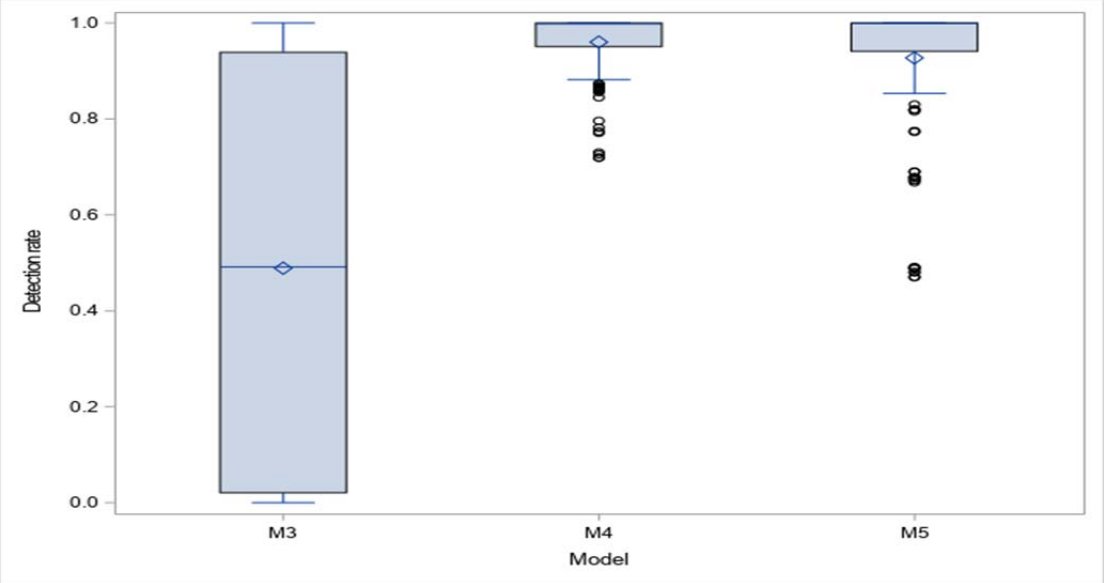


Figure 33. Distributions of detection rates of three models with partially nested data for metric invariance using ΔCFI by type of model

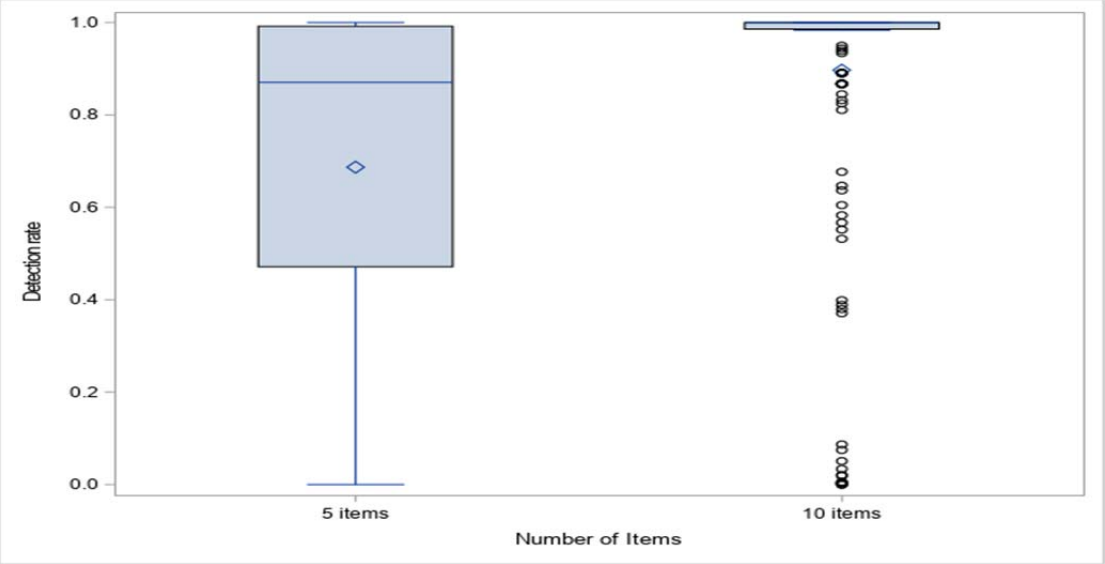


Figure 34. Distributions of detection rates of three models with partially nested data for metric invariance using ΔCFI by number of items

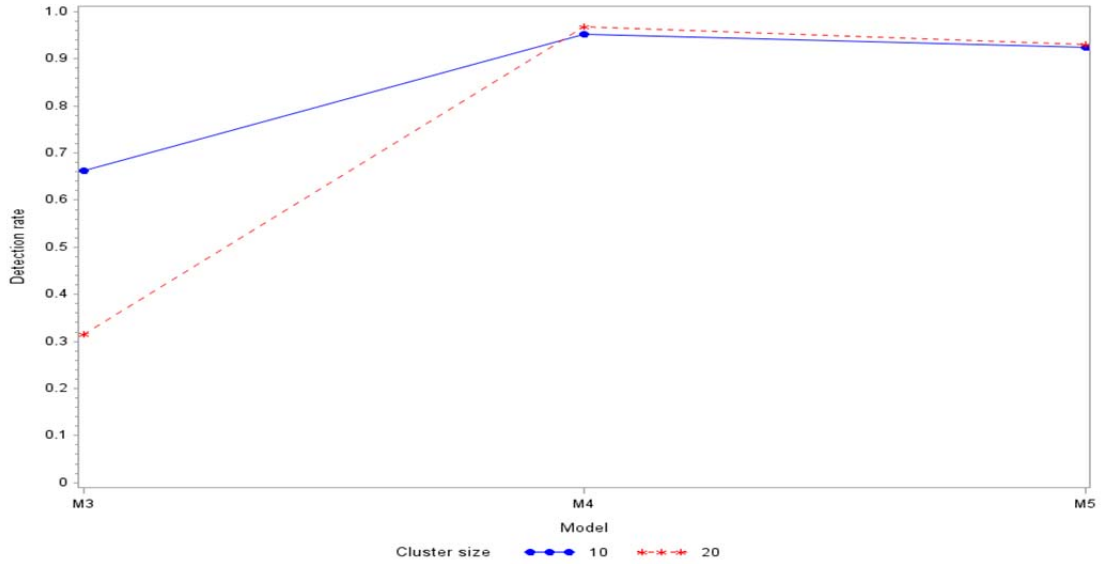


Figure 35. Distributions of detection rates of three models with partially nested data for metric invariance using ΔCFI by model and cluster size

Figure 35 shows that while the detection rates of Model 4 and Model 5 were similar across two levels of cluster size (10 and 20), the detection rates of Model 3 were distinguishable between these two cluster sizes.

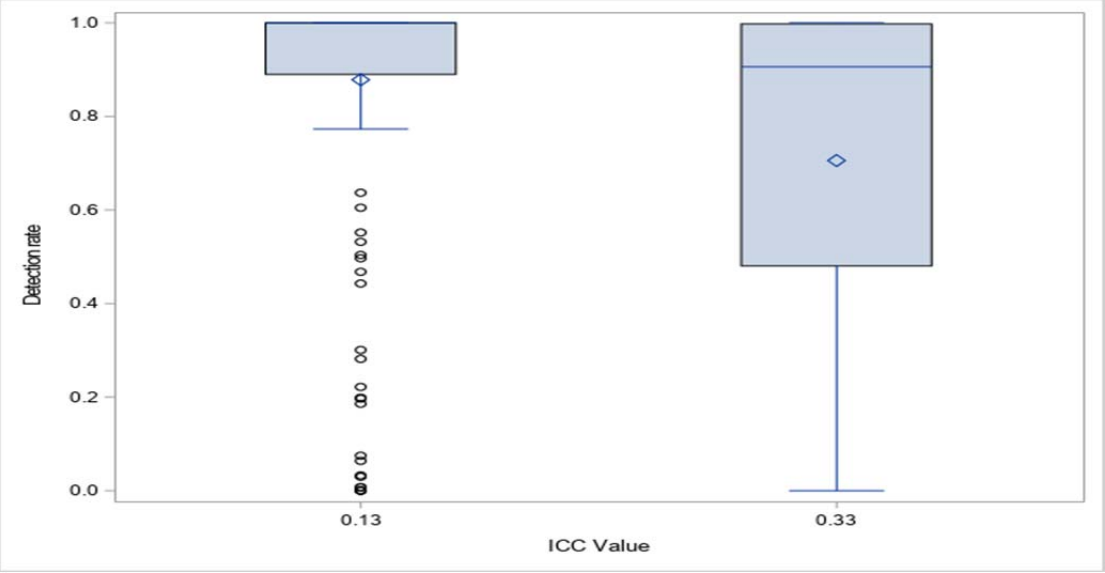


Figure 36. Distributions of detection rates of three models with partially nested data for metric invariance using ΔCFI by ICC

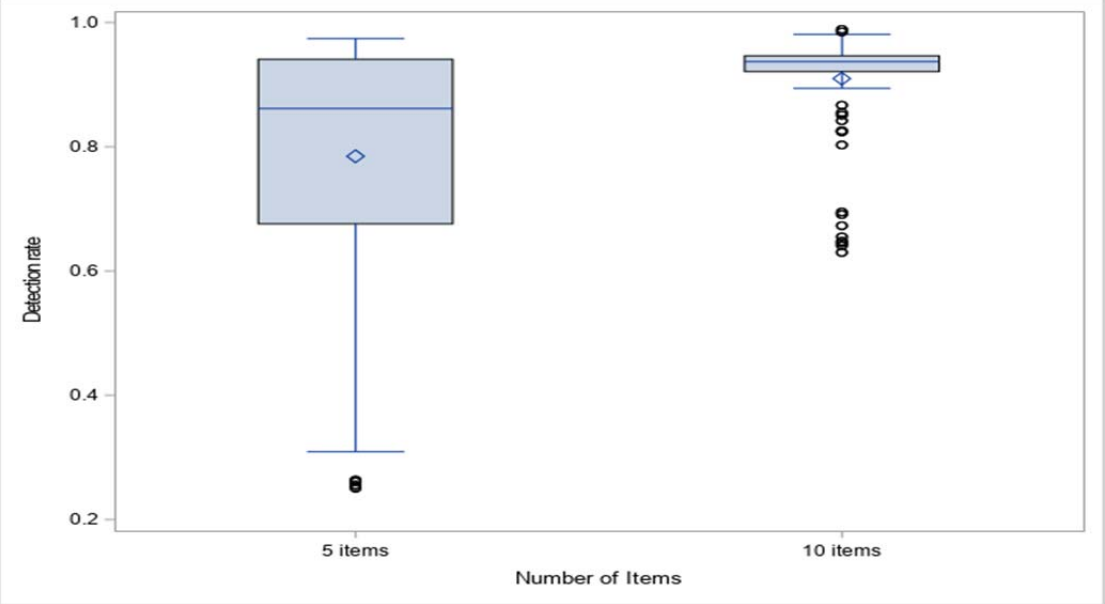


Figure 37. Distributions of detection rates of three models with partially nested data for metric invariance using $\Delta\chi 2$ or SB LRT test by number of items

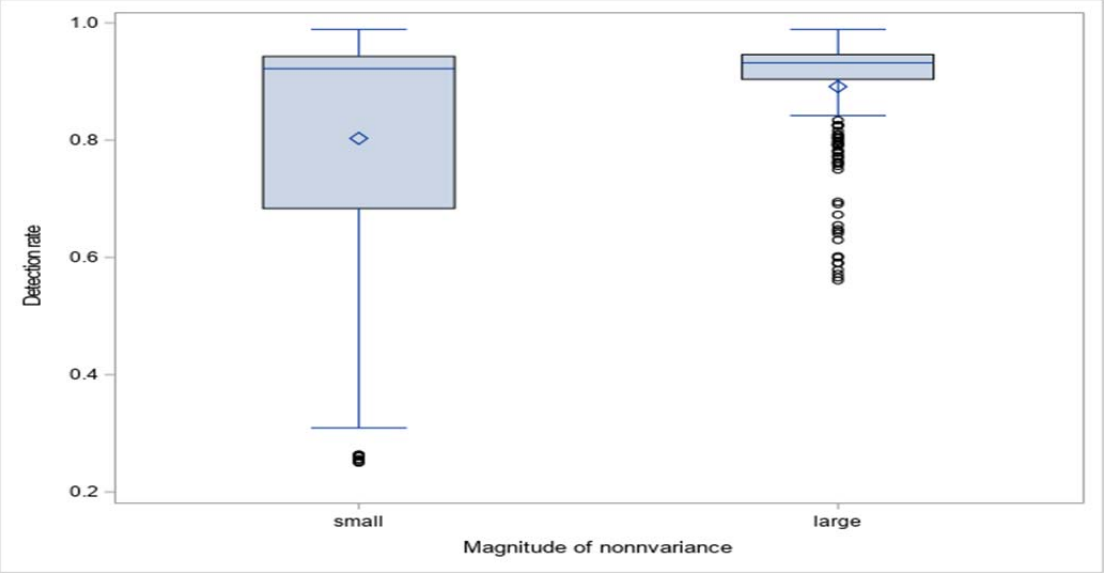


Figure 38. Distributions of detection rates of three models with partially nested data for metric invariance using Δχ2 or SB LRT test by magnitude of noninvariance

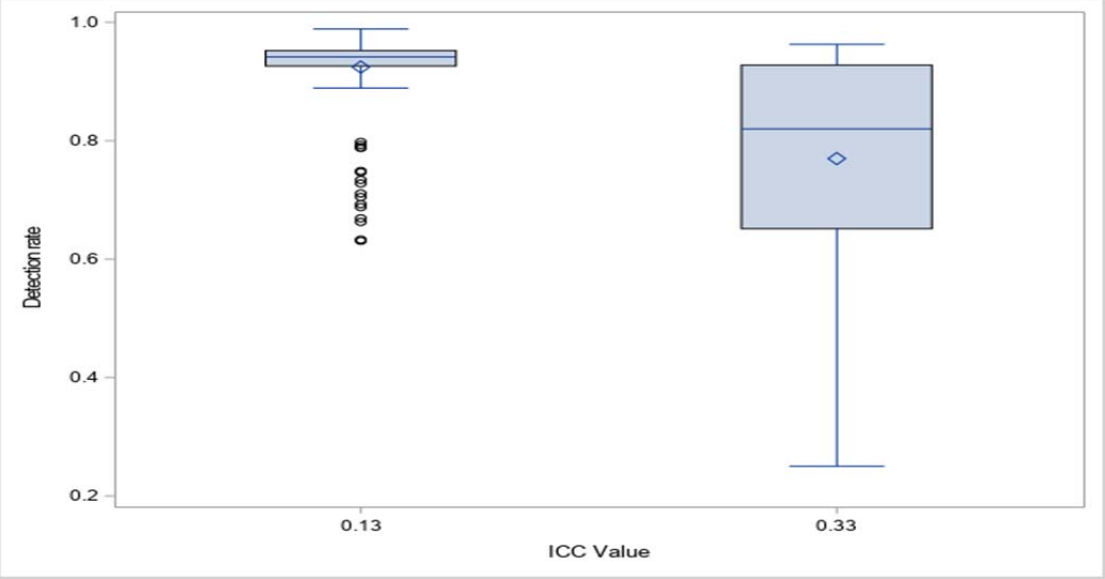


Figure 39. Distributions of detection rates of three models with partially nested data for metric invariance using Δχ2 or SB LRT test by ICC

2.2.3. Effect Sizes for Scalar Invariance Conditions

Table 21 shows the eta-squared values for factors that had significant effects on the detection rates of Models 3, 4 and 5 for scalar invariance conditions. The detection rates of three models were significantly different from each other using all of the three criteria with strongest

effect on the ΔLRT test and least strong on ΔCFI. However the effect of type of model on the detection rates of scalar invariance conditions depended on number of items if using ΔRMSEA (Figure 40) and ICC if using ΔCFI or ΔLRT (Figures 41 and 42). It can be seen in Figure 40 that while there was almost no difference in detection rates of Model 3 between 5-item and 10-item scalar conditions, the difference was noticeable for Model 5 (about 18%) and largest for Model 4 (about 26%).

Table 21. Effect sizes of significant factors on detection rates of Models 3,4, and 5 for scalar invariance

	Model	# of items (i)	ICC (ic)	Number of clusters (c)	Model*ic	Model*i	Model*c
ΔRMSEA	0.348	0.252				0.228	
ΔCFI	0.273		0.151	0.127	0.179		0.122
ΔLRT	0.808				0.069		

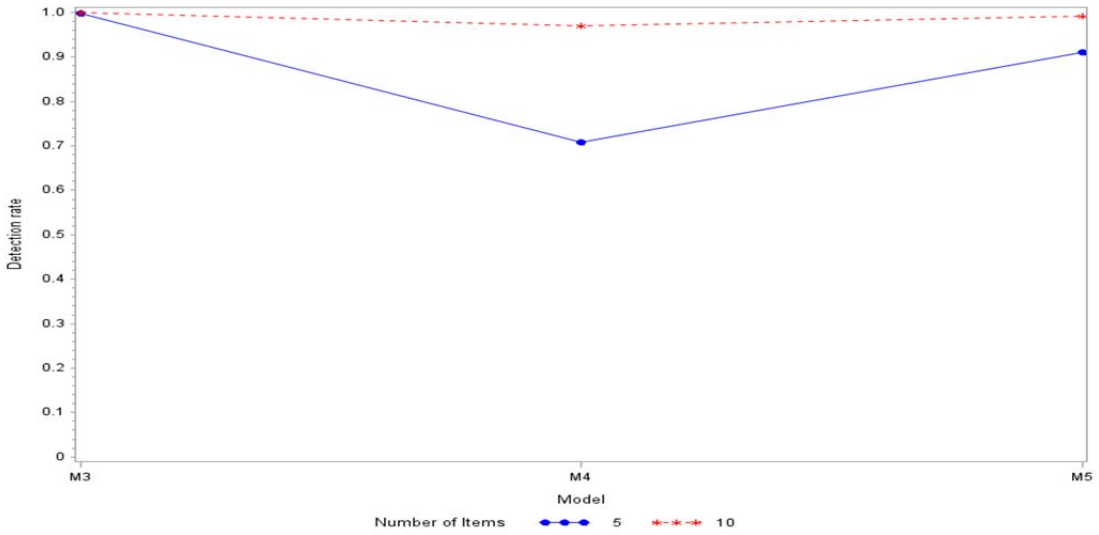


Figure 40. Distributions of detection rates of three models with partially nested data for scalar invariance using ΔRMSEA by type of model and number of items

As witnessed in Figures 41 and 42, the differences in detection rates between two levels of ICC were largest for Model 4 and smallest for Model 3 but the gap was wider if using CFI than using Δχ2 or SB LRT test.

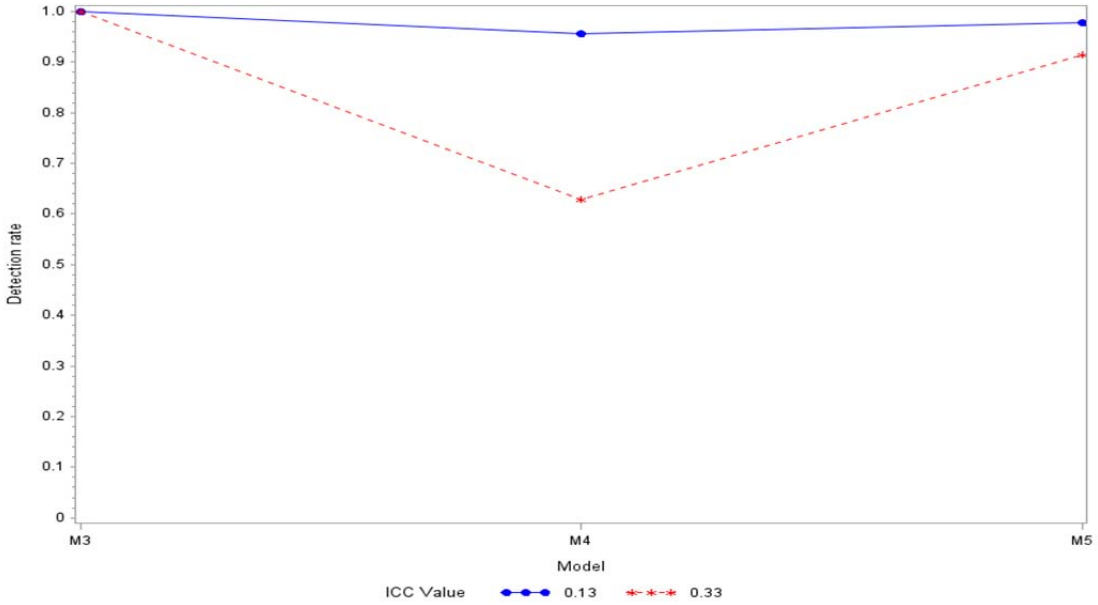


Figure 41. Distributions of detection rates of three models with partially nested data for scalar invariance using ΔCFI by type of model and ICC

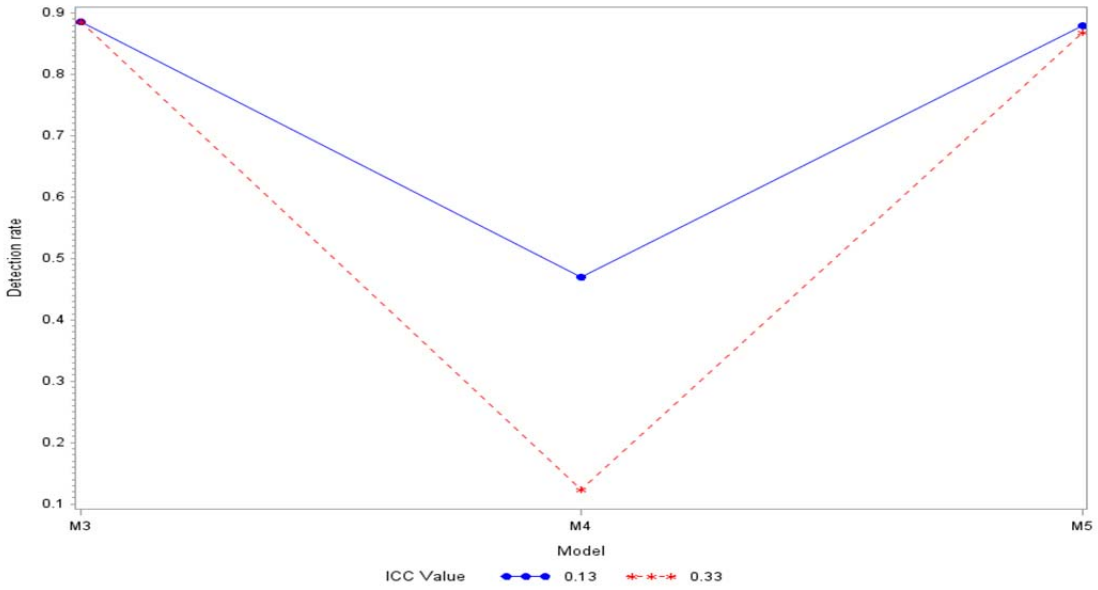


Figure 42. Distributions of detection rates of three models with partially nested data for scalar invariance using $\Delta\chi 2$ or SB LRT test by type of model and ICC

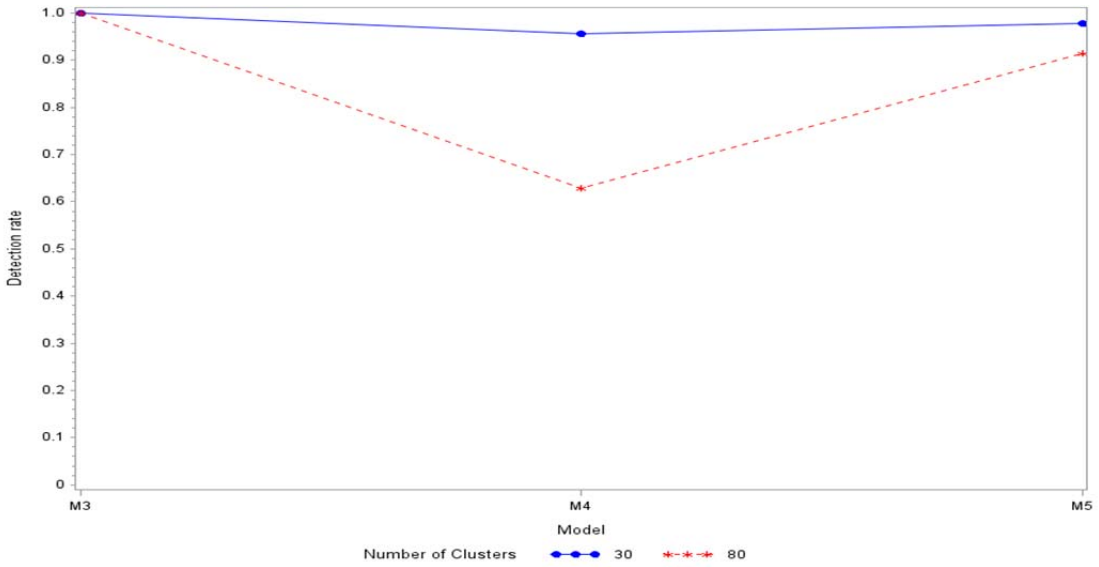


Figure 43. Distributions of detection rates of three models with partially nested data for scalar invariance using ΔCFI by type of model and number of clusters

Similar to the interaction effect of type of model and ICC on the detection rates of scalar invariance using ΔCFI, the differences of detection rates between small number of cluster and large number of cluster were biggest for Model 4, following by Model 5 and smallest (only 1%) for Model 3 as shown in Figure 43.

CHAPTER FIVE: DISCUSSION

1. Summary of the Study

1.1. Purpose

The goals of this dissertation were twofold: (1) propose two statistical models to test measurement invariance between adult informants of children (e.g., father vs. mother, parent vs. teacher, etc.) for paired and partially nested data, and (2) conduct two Monte Carlo simulation studies to investigate the adequacy of the two proposed models as well as the commonly used multiple-group CFA model and the design-based multilevel CFA model.

1.2. Research questions

Research questions for Study 1:

1. How well does each model for the paired data detect the level of measurement invariance (configural, metric, or scalar invariance) under different research settings?

2. What simulation design factors (e.g., factor correlations, degree of data dependency) are related to the performance of the proposed model as well as the comparative model for the paired data?

Research questions for Study 2:

1. How well does each model detect the level of measurement invariance (configural, metric, or scalar invariance) under different research settings for partially nested data?

2. What simulation design factors (e.g., sample size, degree of data dependency) are related to the performance of the proposed model as well as the comparative models for partially nested data?

1.3. Methods

The two studies in this dissertation were Monte Carlo simulations with a partially crossed-factorial design. Study 1 included six simulation factors (number of items, location of measurement noninvariance, magnitude of noninvariance, magnitude of correlation between two informant scores, magnitude of correlation between two unique factors, and sample size). Study 2 was comprised of eight factors (number of items, location of measurement noninvariance, magnitude of noninvariance, magnitude of correlation between two informant scores, magnitude of correlation between two unique factors, number of level-2 units, number of level-1 units per level-2 unit, and partial ICC for nested items).

Mplus 7 software program was utilized to generate data and run the fitted models. The Statistical Analysis System (SAS) package version 9.4 was used to analyze the impact of simulation factors on the outcomes as well as to call and run the fitted models with all replications for each condition in each simulation study. While the estimation method for Model 1, Model 2 and Model 4 was maximum likelihood (ML), the estimator for Model 3 and Model 5 was maximum likelihood estimation with robust standard errors (MLR), which were the default estimators for single level and multilevel or design-based models in Mplus.

2. Answers to Research Questions for Study 1

2.1. Research question 1: How well does each of the two statistical models for the paired data detect the level of measurement invariance (configural, metric, or scalar invariance) under different research settings?

Overall, both Model 1 (repeated measure CFA) and Model 2 (multiple-group CFA) could detect scalar invariance very well with the detection rates from 84% to 100% for Model 1 and slightly higher (85% to 100%) for Model 2 using any of the three criteria. The detection rates of either of the two models in Study 1 using $\Delta\chi 2$ test were pretty similar across all scalar invariance conditions. In other words, no simulation factors examined in Study 1 had significant impact on the detection rates of the two models in Study 1 using $\Delta\chi 2$ test. For these conditions of invariant intercepts and factor loadings, the effects of two simulation factors of magnitude of noninvariance and location of noninvariance on the detection rates did not exist and there was only significant difference in detection rates using $\Delta\chi 2$ test between the two models with higher rates for Model 2 than those of Model 1. However sample size and error correlation had an impact on the detection rates using ΔCFI and ΔRMSEA with higher error correlation or larger sample size resulting in higher detection rates. There was no significant difference between the two models using either of the two global fit indices for scalar invariance conditions.

Unlike the scalar invariance conditions, ability to detect configural invariance or metric invariance was always higher for Model 1 than Model 2 across all cases using any of the three criteria. When sample size was large (500 or 1000), detection rates of Model 1 were always 89% to 100% for configural invariance and 83% or above for metric invariance. These rates for Model 2 were 79% to 100% for configural invariance and 70% to 100% for metric invariance using $\Delta\chi 2$

test. If the magnitude of noninvariance or number of items was large, the detection rates using $\Delta\chi 2$ test were also often high even with a small sample size.

In general, using $\Delta\chi 2$ test or RMSEA difference with cut-off of 0.015 often resulted in higher detection rates of measurement invariance for both Models 1 and 2 than using CFI difference with suggested cut-off of 0.01. CFI difference could result in high detection rates for scalar or metric invariance across most of the research scenarios examined in this study but only led to high detection rates for configural invariance when sample size was 1000 with small noninvariance or sample size of 500 or larger with big noninvariance.

2.2. Research question 2: What simulation design factors (e.g., sample size, degree of data dependency) are related to the performance of the proposed model as well as the comparative models for paired data?

While the magnitude of noninvariance, sample size and the interaction of these two factors all had a significant effect on the detection rates of configural invariance using $\Delta\chi 2$ test or ΔRMSEA criteria for both Models 1 and 2, only magnitude of noninvariance significantly impacted on the detection rates using ΔCFI test. Although higher sample size and magnitude of noninvariance would lead to higher detection rates of configural invariance, the difference in detection rates of small noninvariance coupled with small sample size was much bigger than the difference of large noninvariance coupled with large sample size. In other words, the increase of detection rates of configural invariance using $\Delta\chi 2$ test or ΔRMSEA criteria from small to large noninvariance was slower than the improvement of these rates when sample size increased. The larger the magnitude of noninvariance resulted in much higher the detection rates than the smaller magnitude of noninvariance for the two models for configural invariance while using the ΔCFI test.

Both sample size and number of items were the simulation factors that significantly impacted on the ability to detect the noninvariance in only intercepts and invariant factor loadings (i.e. metric invariance conditions) for the paired data using all of the three criteria: $\Delta\chi 2$, ΔCFI or ΔRMSEA test.

For conditions with invariant factor loadings and intercepts (i.e. scalar invariance), only type of model had eta-squared value larger than significant cutoff of 0.058 on the detection rates using $\Delta\chi 2$ with higher detection rates for Model 2 than Model 1. However the detection rates of scalar invariance for Model 1 were still always higher than 88% across all conditions using $\Delta\chi 2$. Both sample size and number of items were factors that had significant effect on the detection rates using both alternative fit criteria, ΔRMSEA and ΔCFI for the paired data. But the effects of sample size on detection rates using ΔCFI depended on the effect of number of items on detection rates using this criterion.

3. Answers to Research Questions for Study 2

3.1. Research question 1: How well does each of the three statistical models detect the level of measurement invariance (configural, metric, or scalar invariance) under different research settings for partially nested data?

The ability to detect scalar invariance with partially nested data was highest for Model 3, following by Model 5 and lowest for Model 4 for all three criteria. Using any of three criteria, while both Model 3 could detect scalar invariance well with detection rates of nearly 90% across most scalar conditions, Model 4 could only perform this task well for most conditions (except the ones with small number of clusters coupled with large ICC) if using ΔRMSEA or ΔCFI. If using $\Delta\chi 2$ test, Model 4 could moderately detect scalar invariance only with conditions of small cluster size combined with small ICC and could hardly detect scalar invariance for other conditions.

Among the three criteria (i.e., LRT difference, CFI difference, or RMSEA difference), the highest detection rates were often from ΔCFI and ΔRMSEA and the lowest rates were from $\Delta\chi 2$ test. Particularly for Model 3, using ΔCFI and ΔRMSEA always resulted in detection rates of 99% or 100% across all scalar invariance conditions.

Using $\Delta\chi 2$ test or Satorra-Bentler LRT could help detect configural invariance much better for Model 3 than Models 4 and 5, especially for 5-item conditions. With this criterion, while Model 3 could catch configural invariance 82% to 100% among 1000 replication across majority of configural invariance conditions (except only two conditions with detection rates of 61%), Models 4 and 5 could reach these rates only with the largest sample sizes (i.e., large number of cluster or large cluster size or large in both factors) when the magnitude of noninvariance was small. For small sample size conditions combined with small noninvariance, the detection rates for configural invariance were only 31% - 58% for Models 4 and 5. When either magnitude of noninvariance or number of items was bigger, the detection rates for configural invariance were much improved for all of the three models examined in Study 2.

The detection rates for configural invariance were much lower using ΔCFI or ΔRMSEA than $\Delta\chi 2$ test or Satorra-Bentler LRT for Models 3, 4 and 5 with the highest rates for Model 3, following by Model 5 and then Model 4. Probability to detect configural invariance for all three models was only moderate to high using ΔCFI when magnitude of noninvariance was large and largest with combined big noninvariance + large number of items. The detection rates for configural invariance using ΔCFI were very low or low (around 10% - 40%) for all small magnitude of noninvariance conditions. Ability to detect configural invariance using ΔRMSEA was highest for conditions of combined large magnitude of noninvariance + large sample size

(i.e., bigger number of clusters or cluster size or both) for Models 3 and 4 and for conditions of combined large magnitude of noninvariance + small ICC for Model 5.

Unlike scalar and configural invariance, the ability to detect metric invariance was highest for Model 4, following by Model 5 and lowest for Model 3 across many conditions using all of the three performance criteria. The ability to detect metric invariance of Models 3 and 5 using SB LRT test was lowest (25% - 33%) for four conditions of large ICC combined with small noninvariance level + small number of items + small number of clusters. The detection rates of these two models using SB LRT test witnessed much improvement (82% - 97% for the two models) for 10-item conditions, even with small noninvariance or small number of clusters. When using ΔCFI, the detection rates were highest for Model 3 with conditions of either combined small cluster size + small ICC or small cluster size + large noninvariance, highest for Model 5 with all conditions of 10-item (99%-100%) or conditions of 5-item coupled with large magnitude of noninvariance (88% -100%). While the detection rates using ΔCFI were really low (2% - 39%) for small noninvariance + 5-item or small noninvariance + large ICC (0.33) + large number of clusters (80) for Model 3, the overall detection rates using ΔCFI were pretty high to perfect for Models 4 and 5. The detection rates for metric invariance using ΔRMSEA were higher (61% - 100%) and more stable for Model 4 than those of Models 3 and 5 (39% - 100%), especially in comparison with Model 3 (0% - 100%). Model 3 could only detect metric invariance pretty well to perfectly (77% - 100%) in the following conditions: combination of small ICC + large number of clusters + large noninvariance, or small ICC + large number of clusters + small noninvariance + 10-item, or small ICC + large noninvariance + 10-item. In other conditions, Model 3 could never or sometimes catch metric invariance. Models 4 and 5 were able to detect metric invariance very well (mostly above 80% for Model 4 and 95% for

Model 5) in majority of conditions examined in Study 2, except conditions of five items per factor coupled with large ICC.

3.2. Research question 2: What simulation design factors (e.g., sample size, degree of data dependency) are related to the performance of the proposed model as well as the comparative models for partially nested data?

Magnitude of noninvariance was the factor that had strong effect on detection rates using all three criteria for configural invariance and two criteria (ΔLRT and ΔRMSEA) for metric invariance in Study 2. While ICC had a significant effect on detection rates of metric invariance using all three criteria, it only had a significant effect on detection rates of configural invariance using ΔRMSEA. Number of items also had a significant impact on detection rates of the two out three criteria (i.e. ΔRMSEA and ΔLRT for configural invariance, and ΔCFI and ΔLRT for metric invariance conditions) for metric and configural invariance and on detection rates using ΔRMSEA for scalar invariance but in different directions. While larger number of items led to higher detection rates if using ΔLRT or ΔCFI, it resulted in lower detection rates if using ΔRMSEA for both metric and configural invariance. But for scalar invariance conditions, the detection rates of larger number of items conditions were significantly higher than those rates of smaller number of items if using ΔRMSEA.

While type of model had significant effects on detection rates of scalar invariance using any of three criteria, this factor had significant effect on only ΔLRT for configural invariance and on both ΔCFI and ΔRMSEA for metric invariance conditions. Model 3 often had significant higher detection rates for scalar and configural invariance but significant lower detection rates for metric invariance conditions than Models 4 and 5.

4. Discussion and Conclusion

This dissertation examined the performance of two proposed models and other commonly used or potential suitable models to test measurement invariance with paired and partially nested data.

For the paired data, given high detection rates across three levels of measurement invariance (except small sample size conditions in configural invariance), both Model 1 and Model 2 could be reasonable to test measurement invariance with $\Delta\chi 2$ test. However Model 1 could be a favored choice to detect noninvariance due to high and better overall detection rates in metric and configural invariance than those rates of Model 2. The detection rates of metric invariance in Model 1 were always very high across all conditions and those rates of configural invariance were also ranged between 74% and 100% except the four conditions of small sample size combined with small noninvariance. To ensure high probability of detecting all levels of measurement invariance, particularly for configural invariance, sample size of 500 or higher would be recommended. If researchers are interested in testing scalar invariance, Model 2 is a good option with consistently high detection rates for this level of MI.

The results of this dissertation suggested that taking into account the partially nested feature of data (as in Model 3) seems to be more effective in detecting invariance and noninvariance in both factor loadings and intercepts (i.e. scalar invariance and configural invariance, respectively) than detecting noninvariance in intercepts only (i.e. metric invariance).

As higher detection rates for all configural and scalar invariance, and moderate detection rates for many metric invariance conditions (except cases of small number of clusters combined with large ICC), Model 3 could be a good candidate to test measurement invariance with partially nested data. But Model 3 should be used only when having sufficient number of clusters

or if having small number of clusters, the ICC should be also low. Model 5 might be also a reasonable option for this type of data if both the number of clusters and cluster size were large (i.e., 80 and 20, respectively), or either one of these two factors was large coupled with small ICC. If ICC is not small, it is recommended to have a large number of clusters or combination of large number of clusters and large cluster size to ensure high detection rates of measurement invariance for partially nested data. As multiple group CFA had better and reasonable detection rates than the design-based and multilevel repeated measure CFA models cross configural, metric and scalar invariance with the conditions of small cluster size (10) and small ICC (0.13), researchers can consider using this model to test measurement invariance when they can only collect 10 participants within a cluster (e.g. students within a classroom) and there is small degree of data dependency (e.g. small variance between clusters) in the data.

For the paired data, among the configural and metric invariance conditions where the detection rates were low, the next less restricted model in the sequential MI testing was often selected if the correct invariance model was not chosen. Specifically, when the configural invariance was not correctly detected using $\Delta\chi 2$, ΔRMSEA or ΔCFI test, the metric invariance was often selected and the scalar invariance was often chosen in the cases that the metric invariance was not correctly detected.

Based on the research scenario examined in this dissertation, ΔRMSEA and ΔCFI tended to favor scalar invariance when the data were partially nested. When the data were generated as scalar invariance (i.e. invariant factor loadings and intercepts), the detection rates were nearly perfect to perfect using these two criteria. When the data were generated as either metric or configural invariance, if the correct model was not detected, scalar model was also always chosen as the correct model by these two model fit indices regardless the data were generated as

configural or metric invariance. However if using $\Delta\chi 2$ test or SB LRT test, for the incorrect detection cases, if the configural invariance model was not detected, the metric invariance model was chosen more often than the scalar invariance model and if the metric invariance model was not detected, the scalar invariance model was more favorably selected than the metric invariance model.

In general, the larger number of items per factor often led to higher detection rates across three levels of measurement invariance for both studies in this dissertation if using $\Delta\chi 2$ test, SB LRT or ΔCFI. However for partially nested data, while increasing the number of items per factor from five to ten resulted in moderate to very high detection rates for scalar invariance and metric invariance (with highest rates for Model 3), the ability of detecting noninvariance in configural invariance significantly decreased with small number of items if using ΔRMSEA. The ability to detect metric invariance or configural invariance, especially for partially nested data was also poor in many conditions using ΔRMSEA and ΔCFI. This result reiterates the recommendation from Ryu and West (2009) as well as Hsu, Kwok, Lin, and Acosta (2015) about using level-specific global fit indices of ΔRMSEA and ΔCFI for multilevel SEM data due to its lack of power to detect the misspecification in the between-group model. In another aspect, Putnick and Bornstein (2016) suggested some evidence about the relationship of model size (such as degree of freedom, the amount of factors and observed variables estimated in the model) and the performance of fit statistics used in measurement invariance testing (e.g., chi-square difference test, ΔRMSEA, and ΔCFI). However there were lack of studies that examined how model complexity (e.g. number of items per factor, number of parameters estimated) or data complexity (e.g. nested or paired) impacts on the performance of these model fit criteria across different levels of measurement invariance. Further studies are needed to determine how much the

performance of $\Delta\chi 2$ test, ΔRMSEA, and ΔCFI depends on the number of items or other factors in the models used to test different levels of measurement invariance.

5. Limitations of the Study

Given the Monte Carlo research design with control of simulation factors, this dissertation has some limitations. The results of this dissertation are limited to the simulation conditions investigated in the two studies and should be generalized within the extent of these research scenarios. In addition, variables investigated in this dissertation were all continuous with assumption of multivariate normality while categorical variables such as Likert-type scales were also common in educational, social and health sciences.

REFERENCES

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: implications of cross-informant correlations for situational specificity. *Psychological bulletin*, *101*(2), 213.

Achenbach, T. M., & Rescorla, L. (2001). ASEBA school-age forms & profiles.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Asparouhov, T., & Muthén, B. O. (2012). Computing the strictly positive Satorra–Bentler chi-square test in Mplus (Mplus Web Notes No. 12). Retrieved from http://statmodel.com/examples/webnotes/SB5.pdf

Borsboom, D. 2006. When does measurement invariance matter? Commentary. *Medical Care*, 44: S176–S181.

Burns, G. L., Taylor, T. K., & Rusby, J. (2001a). Child and Adolescent Disruptive Behavior Inventory-Version 2.3. *Pullman, WA: Author*.

Burns, G. L., Taylor, T., & Rusby, J. (2001b). Child and Adolescent Disruptive Behavior Inventory (Version 2.3; Teacher). *Pullman, WA: Author*.

Burns, G. L., de Moura, M. A., Walsh, J. A., Desmul, C., Silpakit, C., & Sommers-Flanagan, J. (2008). Invariance and convergent and discriminant validity between mothers' and fathers' ratings of oppositional defiant disorder toward adults, ADHD-HI, ADHD-IN, and academic

competence factors within Brazilian, Thai, and American children. *Psychological Assessment*, *20*(2), 121.

Burns, G. L., & Lee, S .-Y. (2010a). *Child and Adolescent Disruptive Behavior Inventory—Parent Version*. Pullman, WA: Author.

Burns, G. L., & Lee, S .-Y.(2010b). *Child and Adolescent Disruptive Behavior Inventory—Teacher Version*. Pullman, WA: Author.

Burns, G. L., Walsh, J. A., Servera, M., Lorenzo-Seva, U., Cardo, E., & Rodríguez-Fornells, A. (2013). Construct validity of ADHD/ODD rating scales: Recommendations for the evaluation of forthcoming DSM-V ADHD/ODD scales. *Journal of abnormal child psychology*, *41*(1), 15-26.

Burns, G. L., Servera, M., del Mar Bernard, M., Carrillo, J. M., & Geiser, C. (2013). Mothers', fathers', teachers', and aides' ratings of ADHD symptoms and academic impairment: implications for DSM-5 ADHD diagnostic criterion C. *Psychol. Assess*.

Clark, D. A., Listro, C. J., Lo, S. L., Durbin, C. E., Donnellan, M. B., & Neppl, T. K. (2016). Measurement invariance and child temperament: An evaluation of sex and informant differences on the Child Behavior Questionnaire. *Psychological assessment*, *28*(12), 1646.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 464-504.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural equation modeling*, *9*(2), 233-255.

Cohen, J. (1992). Statistical power analysis. *Current directions in psychological science*, *1*(3), 98-101.

De Los Reyes, A., Augenstein, T. M., Wang, M., Thomas, S. A., Drabick, D. A., Burgers, D. E., & Rabinowitz, J. (2015). The validity of the multi-informant approach to assessing child and adolescent mental health. *Psychological Bulletin*, *141*(4), 858.

Dirks, M. A., Reyes, A. D. L., Briggs-Gowan, M., Cella, D., & Wakschlag, L. S. (2012). Embracing not Erasing Contextual Variability in Children's Behavior: Theory and Utility in the Selection and Use of Methods and Informants in Developmental Psychopathology. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *53*(5), 558–574. http://doi.org/10.1111/j.1469-7610.2012.02537.x

De Los Reyes, A., Augenstein, T. M., Wang, M., Thomas, S. A., Drabick, D. A., Burgers, D. E., & Rabinowitz, J. (2015). The validity of the multi-informant approach to assessing child and adolescent mental health. *Psychological Bulletin*, *141*(4), 858.

Duhig, A. M., Renk, K., Epstein, M. K., & Phares, V. (2000). Interparental agreement on internalizing, externalizing, and total behavior problems: A meta□analysis. *Clinical Psychology: Science and Practice*, *7*(4), 435-453.

Fan, X. and Sivo, S. A. 2009. Using Δgoodness-of-fit indexes in assessing mean structure invariance. *Structural Equation Modeling*, 16: 54–69.

Geiser, C., Burns, G. L., & Servera, M. (2014). Testing for measurement invariance and latent mean differences across methods: interesting incremental information from multitrait-multimethod studies. *Frontiers in psychology*, *5*, 1216.

Gregorich, S. E. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework. *Medical care*, *44*(11 Suppl 3), S78.

Gresham, F. M., Elliott, S. N., Cook, C. R., Vance, M. J., & Kettler, R. (2010). Cross-informant agreement for ratings for social skill and problem behavior ratings: An investigation of the Social Skills Improvement System—Rating Scales. *Psychological assessment*, *22*(1), 157.

Harvey, E. A., Fischer, C., Weieneth, J. L., Hurwitz, S. D., & Sayer, A. G. (2013). Predictors of discrepancies between informants' ratings of preschool-aged children's behavior: An examination of ethnicity, child characteristics, and family functioning. *Early childhood research quarterly*, *28*(4), 668-682.

Heck, R. H.,&Thomas, S. L. (2009). *An introduction to multilevel modeling techniques* (2nd ed.). New York, NY: Routledge.

Hox, 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.

Hsu, H. Y., Kwok, O. M., Lin, J. H., & Acosta, S. (2015). Detecting misspecified multilevel structural equation models with common fit indices: A Monte Carlo study. *Multivariate Behavioral Research*, *50*(2), 197-215.

Hsu, H. Y., Lin, J. H., Kwok, O. M., Acosta, S., & Willson, V. (2017). The impact of intraclass correlation on the effectiveness of level-specific fit indices in multilevel structural equation modeling: A Monte Carlo study. *Educational and Psychological Measurement*, *77*(1), 5-31.

Im, M. H., Kim, E. S., Kwok, O. M., Yoon, M., & Willson, V. L. (2016). Impact of Not Addressing Partially Cross-Classified Multilevel Structure in Testing Measurement Invariance: A Monte Carlo Study. *Frontiers in psychology*, *7*, 328.

Jak, S., Oort, F. J., & Dolan, C. V. (2013). A test for cluster bias: Detecting violations of measurement invariance across clusters in multilevel data. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*(2), 265-282.

Kamphaus, R. W., & Reynolds, C. R. (2007). *Behavior Assessment System for Children—Second Edition (BASC–2): Behavioral and Emotional Screening System (BESS)*. Bloomington, MN: Pearson.

Kaplan, D. (2009). *Structural equation modeling: Foundations and extensions* (2nd ed.). Thousands Oaks, CA: Sage.

Kenny, D. A., Veldhuijzen, W., Van Der Weijden, T., LeBlanc, A., Lockyer, J., Légaré, F., & Campbell, C. (2010). Interpersonal perception in the context of doctor–patient relationships: A dyadic analysis of doctor–patient communication. *Social science & medicine*, *70*(5), 763-768.

Kim, E. S. (2011). Testing measurement invariance using MIMIC: Likelihood ratio test and modification indices with a critical value adjustment (Doctoral dissertation). Retrieved from http://oaktrust.library.tamu.edu/bitstream/handle/1969.1/ETD-TAMU-2011-08-9735/KIM-DISSERTATION.pdf

Kim, E. S., & Yoon, M. (2011). Testing measurement invariance: A comparison of multiple-group categorical CFA and IRT. *Structural Equation Modeling*, *18*(2), 212-228.

Kim, E. S., Kwok, O. M., & Yoon, M. (2012). Testing factorial invariance in multilevel data: A Monte Carlo study. *Structural Equation Modeling: A Multidisciplinary Journal*, *19*(2), 250-267.

Kim, E. S., & Cao, C. (2015). Testing group mean differences of latent variables in multilevel data using multiple-group multilevel CFA and multilevel MIMIC modeling. *Multivariate Behavioral Research*, *50*(4), 436-456.

Kim, E. S., Dedrick, R. F., Cao, C., & Ferron, J. M. (2016). Multilevel factor analysis: Reporting guidelines and a review of reporting practices. *Multivariate Behavioral Research*, *51*(6), 881-898.

Kim, E. S., Cao, C., Wang, Y., & Nguyen, D. T. (2017). Measurement invariance testing with many groups: A comparison of five approaches. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(4), 524-544.

Konold, T. R., Walthall, J. C., & Pianta, R. C. (2004). The behavior of child behavior ratings: Measurement structure of the Child Behavior Checklist across time, informants, and child gender. *Behavioral Disorders*, 372-383.

Konold, T. R., & Pianta, R. C. (2007). The influence of informants on ratings of children's behavioral functioning: A latent variable approach. *Journal of Psychoeducational Assessment*, 25(3), 222-236.

Kraemer, H. C., Measelle, J. R., Ablow, J. C., Essex, M. J., Boyce, W. T., & Kupfer, D. J. (2003). A new approach to integrating data from multiple informants in psychiatric assessment and research: Mixing and matching contexts and perspectives. *American Journal of Psychiatry*, 160(9), 1566-1577.

Makransky, G., & Bilenberg, N. (2014). Psychometric properties of the parent and teacher ADHD Rating Scale (ADHD-RS) measurement invariance across gender, age, and informant. *Assessment*, *21*(6), 694-705.

Mayfield, A. R., Parke, E. M., Barchard, K. A., Zenisek, R. P., Thaler, N. S., Etcoff, L. M., & Allen, D. N. (2018). Equivalence of mother and father ratings of ADHD in children. *Child neuropsychology*, *24*(2), 166-183.

McCarthy, K. E. S. (2015). *Predicting achievement in children: A comparison of methods and raters* (Doctoral dissertation). Retrieved from https://search.proquest.com/docview/1690497489?pq-origsite=gscholar

Meade, A. W. and Bauer, D. J. (2007). Power and precision in confirmatory factor analytic tests of measurement invariance. *Structural Equation Modeling*, 14: 611–635.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research, 13*(2), 127-143.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525-43.

Meredith, W. and Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical Care*, 44: S69–S77. Millsap, R. E. and Kwok, O.-M. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Methods*, 9: 93–115.

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., ... & Reed, G. M. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American psychologist*, *56*(2), 128.

Multilevel Modeling Tutorial Using SAS, Stata, HLM, R, SPSS and Mplus. Retrieved from https://stat.utexas.edu/images/SSC/documents/SoftwareTutorials/MultilevelModeling.pdf

Muthén, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. Sociological Methodology, 25, 267–316.

Muthén, L.K. & Muthén, B.O. (1998-2010). Mplus User's Guide. Sixth Edition. Los Angeles, CA: Muthén & Muthén.

Muthén, B., & Asparouhov, T. (2002). Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus. *Mplus web notes*, *4*(5), 1-22.

Narad, M. E., Garner, A. A., Peugh, J. L., Tamm, L., Antonini, T. N., Kingery, K. M., ... & Epstein, J. N. (2015). Parent–teacher agreement on ADHD symptoms across development. *Psychological Assessment*, *27*(1), 239.

Olsen, J. A., & Kenny, D. A. (2006). Structural equation modeling with interchangeable dyads. *Psychological methods*, *11*(2), 127.

Piskernik, B., Supper, B., & Ahnert, L. (2018). Measurement Invariance Analysis of the Parental Stress Index. *European Journal of Psychological Assessment.*

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental review, 41,* 71-90.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods, Second Edition.* Newbury Park, CA: Sage.

Renk, K., & Phares, V. (2004). Cross-informant ratings of social competence in children and adolescents. *Clinical Psychology Review*, *24*(2), 239-254.

Robey, R. R., & Barcikowski, R. S. (1992). Type I error and the number of iterations in Monte Carlo studies of robustness. *British Journal of Mathematical and Statistical Psychology*, *45*(2), 283-288.

Rothbart, M. K., Ahadi, S. A., Hershey, K. L., & Fisher, P. (2001). Investigations of temperament at three to seven years: The Children's Behavior Questionnaire. *Child development*, *72*(5), 1394-1408.

Ryu, E., & West, S. G. (2009). Level-specific evaluation of model fit in multilevel structural equation modeling. *Structural Equation Modeling*, *16*(4), 583-601.

Ryu, E. (2014). Factorial invariance in multilevel confirmatory factor analysis. *British Journal of Mathematical and Statistical Psychology*, *67*(1), 172-194.

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis.

Satorra, A., & Bentler, P. M. (2010). Ensuring positiveness of the scaled chi-square test statistic. Psychometrika, 75, 243–248. doi:10.1007/ s11336-009-9135-y

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review*, 18, 210–222.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling (2nd ed.).* Thousand Oaks, CA: Sage

Sterba, S. K., Preacher, K. J., Forehand, R., Hardcastle, E. J., Cole, D. A., & Compas, B. E. (2014). Structural equation modeling approaches for analyzing partially nested data. *Multivariate Behavioral Research, 49*(2), 93-118.

Tröster, H. (2011). Eltern-Belastungs-Inventar. Deutsche Version des Parenting Stress Index (PSI) von R. R. Abidin [German version of the Parenting Stress Index (PSI) by R. R. Abidin]. Goettingen, Germany: Hogrefe.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4-69.

Van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, *9*(4), 486-492.

Waschbusch, D. A., & Willoughby, M. T. (2008). Parent and teacher ratings on the IOWA Conners Rating Scale. *Journal of Psychopathology and Behavioral Assessment*, *30*(3), 180.

Woehr, D. J., Sheehan, M. K., & Bennett Jr, W. (2005). Assessing measurement equivalence across rating sources: a multitrait-multirater approach. *Journal of Applied Psychology*, *90*(3), 592. doi:10.1037/0021-9010.90.3.592

Woods, C. M. (2009). Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. *Multivariate Behavioral Research, 44*(1), 1-27.

Yoon, M. and Millsap, R. E. (2007). Detecting violations of factorial invariance using data-based specification searches: A Monte Carlo study. *Structural Equation Modeling*, 14: 435–463.

Yoon, M. (2008). Statistical power in testing factorial invariance with ordinal measures. *Dissertation Abstracts International, 68*(11), 7705B-7854B.