Summer 2022

Investigating Performance of Model Fit Indices In Multiple-Group Confirmatory Factor Analysis: Complications With Ordinal Data

Ning Jiang

INVESTIGATING PERFORMANCE OF MODEL FIT INDICES IN MULTIPLE-GROUP CONFIRMATORY FACTOR ANALYSIS: COMPLICATIONS WITH ORDINAL DATA

by

Ning Jiang

Bachelor of Arts
Dalian University of Foreign Languages, 2012

Master of Education
University of South Carolina, 2016

Submitted in Partial Fulfillment of the Requirements

For the Degree of Doctor of Philosophy in

Educational Psychology and Research

College of Education

University of South Carolina

2022

Accepted by:

Christine DiStefano, Major Professor

Tammiee Dickenson, Committee Member

Dexin Shi, Committee Member

Jin Liu, Committee Member

Tracey L. Weldon, Vice Provost and Dean of the Graduate School

DEDICATION

I would like to dedicate my dissertation work to my parents, Xingtai Jiang and Hui Jin, who have provided and continue to provide unconditional love to me. I also dedicate this dissertation to my husband, Junliang Liu, who has been a constant source of support and encouragement throughout my graduate school and life. Thank you for listening to my worries about the dissertation and believing me throughout the process. Last, this work is also dedicated to my first child, Felix Jiang Liu. Thank you for coming to this family and choosing me as your mother. You are a special gift for this journey. There are no words to express my gratitude, but I hope this work makes all of you proud.

ACKNOWLEDGEMENTS

At the age of 24, I graduated from college, left my hometown, and started my study journey to the United States. If someone asked me why I chose the University of South Carolina at Columbia, I could only say that it is a university located in a southern city, where the warm winter is enjoyable. Coming here, I had nothing but was full of curiosity about my major, professors, and classmates. I still remember the difficulties that I faced and the tears that I shed in my first year, but I confirmedly told myself: Hey girl, your dream will come true only if you work hard.

Seven years later, I am still living here, but I no longer have nothing. I have professors, friends, colleagues, my love, and my little son. If someone asked me again why I chose this place, I would tell them, because it has memories of my struggles and happiness, also, because of my professors, colleagues, and friends who make me feel at home. Without their encouragement and help, I would not have completed my program, and without their company and support, I would not have loved this city. Along the way, there are so many people that I would like to express my gratitude to.

First, I would like to express my deepest gratitude to my mentor, advisor, and role model, Dr. Christine DiStefano. Thank you for bringing me into the academic field of educational research. I knew almost nothing about this field in my first year of this program. However, you did not give up on me. Thank you for believing me and being willing to be my mentor! You are like a mother teaching a child to walk, guiding me

patiently, and giving me invaluable support. I am so grateful for everything you did for me. Hopefully, someday, you will be proud of me.

Second, I would like to acknowledge my dissertation committee members, Dr. Tammiee Dickenson, Jin Liu, and Dexin Shi. I would like to thank you for your support and feedback. I am very grateful to Dr. Dickenson for allowing me to work as a research assistant in the Research, Evaluation, and Measurement Center. I am also grateful to Dr. Liu, you are not only my academic leader but also my big sister in life. I am especially grateful to Dr. Shi, who gave me the initial inspiration for the dissertation topic, and your suggestions helped me a lot. In addition, I would also like to thank Dr. Maydu-Olivares for your advice that makes my dissertation comprehensive.

Third, I would like to give special thanks to my project director Dr. Ashlee Lewis, and my colleagues: Bradley Rogers, Ruiqin Gao, Xumei Fan, Jiali Zheng, and Bryanna Montpeirous. Your trust and continuous support make me believe that I can do my job well. Dr. Lewis, I especially thank you for giving me many work opportunities and allowing me to join various projects to learn, explore, and practice. It is a great pleasure to work with you.

Finally, I want to thank my family. Without your unconditional love, I would not have started my American Dream.

ABSTRACT

The purpose of this study is to evaluate the performance of three commonly used model fit indices when measurement invariance is tested in the context of multiple-group CFA with categorical-ordered data. As applied researchers are increasingly aware of the importance of testing measurement invariance, as well as Likert-type scales are frequently used in the social and behavioral sciences, specific guidelines are in need for establishing measurement invariance using model fit indices.

To achieve the study goal, two Monte Carlo simulation studies were conducted. Study 1 investigated the sampling variability of fit indices under different levels of invariance tests. Based on the sampling variability of fit indices, cutoff values for various levels of invariance were proposed. Study 2 investigated the influence of several conditions on the sensitivity of changes in fit indices to two commonly used non-invariance levels: metric non-invariance and scalar non-invariance. Then, rejection rates based on cutoff values of proposed fit indices were examined in Study 2.

Findings indicated that all three fit indices (CFI, RMSEA, and SRMR) appeared to be more sensitive to lack of invariance in thresholds than loadings. Different cutoff values may be applied under various conditions with categorial-ordered data. In addition, cutoff values should be used with caution as factors impacted changes in model fit indices differently. Recommendations for the use of model fit indices in the multiple-group CFA invariance context were provided for applied researchers.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

In the social and behavioral sciences, researchers use various instruments (e.g., surveys, tests, questionnaires) to collect data for use in investigating characteristics of latent constructs. When researchers use these instruments, ensuring the validity associated with the factor scores is a major issue in psychometrics. More simply, the measures should produce precise measurements of the concepts they are supposed to be measuring (Bauer, 2017), and as such, provide evidence to assist in the interpretation of the underlying latent constructs.

Validity evidence takes many forms including face validity, content validity, criterion validity, and construct validity (Rawls, 2009). Considering construct validity, one area which may be of importance to researchers is to examine the equivalence of a latent construct across different conditions (e.g., gender, ethnicity, or occasions of measurement). By assessing invariance, one can ensure that the focal construct is measured and interpreted in the same manner across groups.

In the educational literature, the procedure of examining the equivalency of the construct of interest across groups is called measurement invariance (MI). If researchers are interested in making comparisons across groups or measurement occasions, measurement invariance must be satisfied. This will ensure that the indicator responses depend only on latent scores and not on the group membership. In this way, the

differences in the observed scores can accurately reflect the true (i.e., latent) differences of the construct being measured. When measurement invariance does not hold, the measure produces scores that differ among groups due to "other" factors rather than differences on the latent variable. For example, scores on a measure may be more strongly endorsed by one group than another after controlling for the latent construct of interest. Instead of measuring only individual differences on the focal construct, we are also measuring irrelevant factors (Meredith, 1993). As a result, the measure fails to accurately reflect true differences of the targeted construct and making group comparisons is questionable. Thus, the evaluation of measurement invariance is an essential step for researchers to be able to draw valid conclusions about latent construct differences across groups or measurement occasions.

Researchers are becoming increasingly aware the importance of testing measurement invariance. The literature on this topic has rapidly increased since 1990 (Bauer, 2016). Specifically, in a search of the APA PsycInfo database between 1974-2020 with the exact phrase in the title and abstract- Measurement Invariance, only 14 published articles related to measurement invariance were found before 1990 (see Figure 1.1). However, after 1990, the number of MI articles increased dramatically to 3,855, with over 85% of these studies (2,831) conducted in the last 10 years (2010- 2020).

Additionally, a substantial number of studies related to measurement invariance have been published across many applied disciplines such as education, psychology, developmental psychology, marketing, and organizational sciences (Vandenberg & Lance, 2000). This substantial increase in the use of measurement invariance relies upon

the rapid developments in statistics, which provided many new analytical tools for assessing measurement invariance in applied contexts.

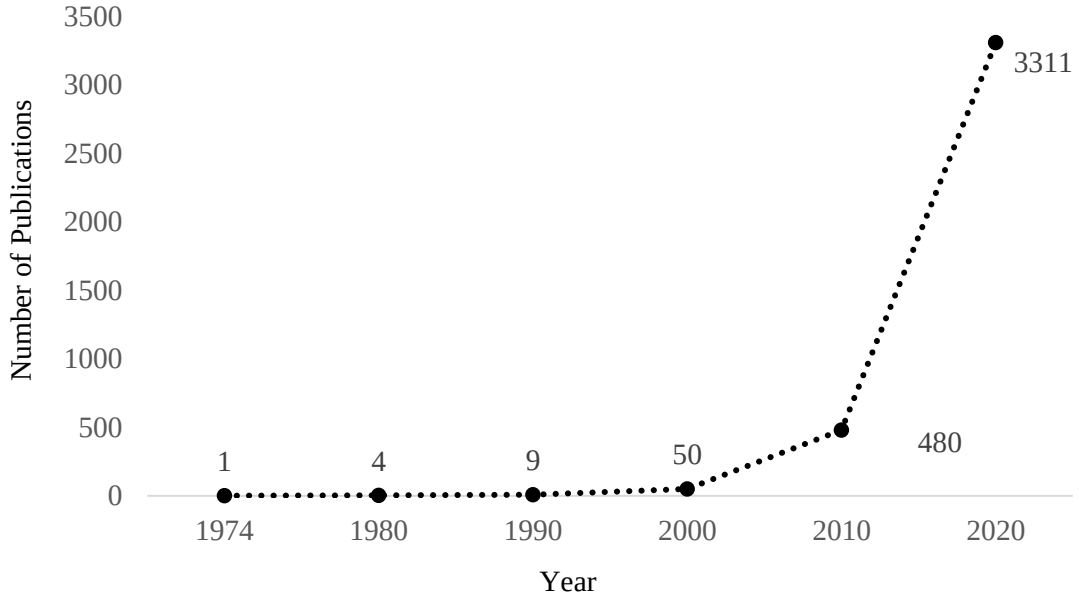

Figure 1.1 The trend of publications related to measurement invariance from 1974 to 2020. Note: Data are from APA PsycInfo database.

Measurement invariance is often detected by using latent variable modeling techniques, such as structural equation modeling (SEM) or item response theory (IRT). Confirmatory factor analysis (CFA) is the most popular analytical strategy to explore the underlying latent structure among a set of observed variables and to provide evidence of construct validity in theory-based instrument construction (Li, 2016). CFA was originally developed for use with continuous indicators, and the maximum likelihood (ML) estimation method is often employed to estimate model parameters. When ML is used to estimate CFA model parameters, observed data need to follow the assumption of multivariate normality. Given multivariate normally distributed data, adequate sample size, and proper model specification, ML provides consistent, efficient, and unbiased parameter estimates and asymptotic standard errors (Bollen, 1989).

However, one major consideration of examining CFA models is that the indicators or survey items in the social and behavioral sciences are often ordinal (i.e., Likert-type scale items) and the observed data often present some degree of non-normality (DiStefano & Morgan, 2014). If a questionnaire uses a Likert scale to collect data, the use of the ML estimator is no longer appropriate, as the multivariate normality assumption does not hold. Further, it is inappropriate to use the ML estimator when data only have a few response categories because the multivariate normality is severely violated, exhibiting high levels of skewness and/or kurtosis (DiStefano & Morgan, 2014; Lubke & Muthén, 2004).

To accommodate the non-normal nature of ordinal data, several alternative estimation techniques have been developed to evaluating the hypothesized relations among ordinal variables, such as diagonally weighted least squares (DWLS), weighed least squares (WLS), or robust weighted least squares (WLSMV). Prior research has noted that these alternative estimators may perform better than the ML estimator when ordinal data are analyzed (e.g., Flora & Curran, 2004; Sass, 2011; Yuan & Bentler, 2000). For example, factor loading estimates by WLS and WLSMV were less biased than ML estimator (Beauducel & Herzberg, 2006); and comparing to ML-based chi-square values, WLSMV-based chi-square values exhibited a lower Type I error rate when categories were small (Beauducel & Herzberg, 2006). Flora and Curran (2004) concluded that overall performance of WLSMV was superior to the performance of WLS across almost all conditions. Generally, previous studies have documented that WLSMV estimator for categorical data performs better across many conditions often encountered in empirical work (DiStefano & Morgan; 2014; Schmitt, 2011). Given the benefits, WLSMV is

recommended for estimating CFA model parameters when data are ordered-categorical (Muthén & Muthén, 2010).

Within the CFA framework, one common approach to assess measurement invariance is the multiple-group (MG) model. When using the multiple-group CFA (MG-CFA) framework to evaluate across-group invariance, researchers often divide the data by groups, and different confirmatory factor models are compared to identify similarities and differences across groups by imposing equality constraints on model parameters (e.g., item loadings, intercepts, or residuals). Based on degrees of model parameter constraints, measurement invariance levels are often reported hierarchically. Vandenberg and Lance (2000) illustrated eight levels of measurement invariance: covariance matrices (invariance of covariance matrices across groups), configural (invariance of the model form), metric (invariance of item loadings), scalar (invariance of item intercepts), strict (residual invariance), as well as equivalence of factor variances, covariances, and means.

To determine invariance level of CFA models across groups, most researchers rely on assessing model fit statistics. The likelihood ratio difference test (differences in chi-square) has been the most frequently used statistic for testing the difference between nested models (i.e., a baseline model vs. a more restricted model) (Chen, 2007; Putnick, 2016). If the result of likelihood ratio difference test indicates non-significance, then the model with more restricted parameter constraints performs as well as the baseline model. Then, further constraints on parameters can be added to test a higher level of invariance.

Likelihood ratio difference test, however, are with several limitations. First, it is sensitive to sample size (Chen, 2007). Specifically, studies have shown that both small and large sample sizes may impact chi-square results, which may lead to the false

rejection of models (Bergh, 2015), and induce bias of parameter or standard error estimates (Flora & Curran, 2004; Babyak & Green, 2010). Second, although the likelihood ratio statistic can be estimated using various estimation methods, the maximum likelihood (ML) is the most popular estimator used by applied researchers (Hu & Bentler, 1999). Thus, it is important for researchers to make sure data follows the multivariate normality (Alavi et al., 2020). Third, instead of testing whether measurement invariance holds approximately, likelihood ratio test assesses whether measurement invariance holds exactly, which may be too stringent for practical purposes (Ene, 2020).

In such cases, alternative model fit indices, such as the Comparative Fit Index (CFI), the Tucker-Lewis Index (TLI), the Root Mean Squared Error of Approximation (RMSEA), or the Standardized Root Mean Square Residual (SRMR) are recommended as a supplement to evaluate measurement invariance. Model fit is used to evaluate how well the proposed model fits a set of data. With invariance testing, configural invariance is tested by evaluating multiple fit indices using common benchmarks. To determine the rest levels of invariance, model fit indices of two nested models are compared. Researchers can decide whether the more restricted model, with more model parameters constraints imposed, fits less well than the less restricted model with fewer constraints on the model parameters (Vandenberg & Lance, 2000). For example, between the nested models of configural and metric models, the only difference between the two models is attributed to the imposed constraints on the factor loadings. By computing the difference between fit statistics for two models (e.g., ΔCFI, ΔRMSEA), researchers can decide whether metric invariance holds or not comparing to the configural invariance.

Rather than using "typical" cutoffs, several researchers have proposed criteria to evaluate changes in fit indices when testing measurement invariance with continuous data. For example, the commonly used criteria in applied studies are Cheung and Rensvold's (2002) criterion of a -0.01 change in CFI between nested models, as well as Chen's (2007) criterion of a -0.01 change in CFI, combined with changes in RMSEA of 0.015 and SRMR of 0.03 (for metric invariance) or 0.015 (for scalar or residual invariance). However, in the measurement invariance literature, only a few research studies have investigated criteria of differences in model fit indices (e.g., ΔRMSEA, Δ CFI, ΔSRMR) (e.g., Chen, 2007; Cheun, & Rensvold, 2002; Rutkowski & Svetina, 2014; Rutkowski, & Svetina, 2017). The research is even more limited when considering how to evaluate changes in fit when analyzing ordered-categorical data. Although the use of measurement invariance with multiple-group CFA in applied research has rapidly increased, there are not generally criteria in MG-CFA for determining invariance based on changes in model fit indices when ordered-categorical data are analyzed.

Given the importance of measurement invariance in the social and behavioral sciences, there is a gap in the literature concerning changes of model fit indices when ordered-categorical data are analyzed with MG-CFA. This study filled the existing gap by examining the performance of model fit indices under a broad range of conditions when ordinal data are analyzed. Specifically, the current study aimed to answer the following research questions:

1) What criteria of changes in model fit indices should be recommended to evaluate measurement invariance of ordinal data?

2) Are proposed criteria of changes in model fit indices consistent with measurement invariance at the different levels: factor loadings, thresholds, residual variances, factor variance, covariance, and mean?
3) What conditions impact the performance of changes in fit indices on detecting the measurement invariance?
4) Are currently proposed standards for evaluating measurement invariance with continuous data suitable for the measurement invariance with ordinal data?

CHAPTER 2

LITERATURE REVIEW

Measurement invariance is an important aspect of instrument validation, and the establishment of measurement invariance is considered a prerequisite to examining group differences (i.e., invariance of model parameter estimates). When group differences occur based on the constructs that instruments tend to measure, the conclusions drawn from a study may be inaccurate as the constructs may not reflect the same meanings across different groups. In other words, inferences about group differences may be due to measurement bias and not due to different positions on the latent variable(s). Although measurement invariance has been examined in many previous studies, the investigation of criteria to use when examining model fit is limited. Applications of fit indices with ordered categorical data under the measurement invariance topic are even more sparse. The purpose of this study was to examine model fit indices commonly used for evaluation of measurement invariance when ordered categorical data are analyzed.

This chapter presented an overview of literature on measurement invariance testing in social sciences. I started with an introduction to the history and statistical modeling development of measurement invariance testing. Definitions on different types of measurement invariance were provided. A review of invariance testing using multiple-group confirmatory factor analysis (MG-CFA) with both continuous and ordinal data was also given. After that, I focused on an overview of steps involved in conducting the

multiple-group CFA, including invariance constraints, estimation methods, and fit evaluation. The chapter ended up with a summary of the current gaps in the research.

2.1 Measurement Invariance

The idea of testing measurement invariance (MI) can be traced back approximately 60 years ago, when the public started to be concerned that cognitive ability tests were unfair to minority examinees (e.g., Angoff, 1993; Meredith, 1964; Walker, 2011). Since that time, researchers have emphasized the importance of establishing measurement invariance for test validation, and considerable discussion has been devoted to whether the latent construct of an instrument (e.g., ability, depression) is measured in the same way when it is administered across groups such as gender, ethnicity, culture, or across time points. Without establishing measurement invariance, any observed differences (e.g., means, regression coefficients) across groups/time may not reflect the true differences in the constructs of interest (Shi, 2016). Under such situations, making group comparisons is questionable as violations of measurement invariance threaten the reliability and validity associated with construct scores. Much research has been done on this issue, and with the rapid development of new statistical techniques in recent decades, researchers in psychology and education are able to use advanced statistical approaches to investigate measurement invariance, especially within the latent variable modeling frameworks (e.g., Cheung & Rensvold, 2002; Mellenbergh, 1989; Millsap, 2011; Rensvold & Cheung, 1998; Widaman & Reise, 1997).

Measurement invariance is a general term that can be applied to a variety of psychometric models. In psychological and educational research, measurement invariance is often studied within latent variable modeling frameworks such as SEM,

IRT, and some methodologists have even suggested integrating the two approaches (e.g., Reise, et al., 1993; Widaman & Grimm, 2014). Among these approaches, multiple-group confirmatory factor analysis (MG-CFA) (Jöreskog, 1971; Meredith, 1993) is the most widely used method in applied research. In general, this analytic approach tests whether the linear relationship between the latent factor and observed variables is consistent with the hypothesized model across multiple groups by imposing constraints on the model parameters of the nested CFA models. Researchers typically assess various levels of measurement equivalency by comparing a series of model fit statistics. Prior studies have shown that this modeling strategy is very adaptable to determine the equivalence of an instrument's psychometric properties at both item and latent levels (Sass, 2011). In addition, it has many potential practical applications such as evaluation of psychometric scale development (Bagozzi & Edwards, 1998), detection of item bias (Woods & Grimm, 2011), assessment of longitudinal change (Putnick & Bornstein, 2016), or cross-group comparisons (Brown, 2015).

Statistical approaches to detect measurement invariance under the IRT framework are also popular in large-scale high-stakes educational testing programs. Within the IRT context, MI is evaluated by examining whether the multi-items related to the construct(s) perform the same for all individuals. If the relationships exhibit differences, then differential item functioning (DIF) is present. DIF is defined as "the circumstance in which two individuals of similar ability do not have the same probability of answering a question in a particular way" (VandenBos, 2014; p.93). Since this study focused on measurement invariance using multiple-group confirmatory factor analysis, invariance testing approaches under the IRT framework are not described in detail. Readers

interested in measurement invariance under the IRT framework can refer to Tay, et al. (2015) and Millsap (2011) for an overview.

2.1.1 Definition of Measurement Invariance

Researchers have provided definitions of measurement invariance. According to Mellenbergh (1989) and Millsap (2011), measurement invariance can be expressed mathematically as:

$$f_1(Y_i|T = t,\ V_i = v_i) = f_2(y_j|T = t)$$

Where Y_i is a $p \times 1$ vector containing the observed items for the person i, T is the latent construct that is measured by Y_i, and V_i is a $q \times 1$ vector of a set of conditions (e.g., gender, ethnicity, age, occasions, or test settings). The function f_1 represents the conditional distribution function of Y_i given t and v, and f_2 is the conditional distribution function of Y_i given t. This equation states that the conditions (V_i) do not directly influence the distribution of observed scores (Y_i) other than through the influence on the underlying latent variable, T (Bauer, 2016).

If $f_1 \neq f_2$, meaning conditional independence does not hold, then it is stated that an item lacks MI (or exhibits DIF) in relation to V_i. In other words, the measurement of T by Y_i is said to be biased with respect to V_i. If $f_1 = f_2$, this indicates that MI holds, and the distribution of the observed items only depends on the values of the latent traits or latent variables without bias.

To better understand the above mathematical formula, several points should be noted. First, the above mathematical definition permits us to define measurement bias as a violation of measurement invariance (Millsap, 2011). Second, measurement bias/non-invariance is referred to as systematic inaccuracy in measurement (Millsap, 2011). In

contrast to random errors of measurement, the inaccuracy in measurement is replicable. For example, a test item may demonstrate that male students have a higher mean score on an item than female students, and this finding is observed across multiple samples of male and female students. Then, we may conclude that the higher mean score on an item for males and females is not a result of random sampling errors. Third, according to Millsap, researchers should differentiate the concept of systematic inaccuracy from systematic group differences. Systematic inaccuracy only occurs if the item does not reflect the actual status of the construct being measured (Millsap, 2011). For example, if male students truly have a higher ability than female students, the test should produce different scores by gender. However, if male students and female students exhibit score differences by gender, we may have two kinds of ambiguous interpretations: the gender difference could reflect a real difference in the latent construct of ability, or the item is systematically inaccurate (or is biased) in relation to gender.

Mellenbergh (1989) clarified the concept of measurement invariance by bringing a matching criterion into the definition. Specifically, the concept of measurement bias/non-invariance is based on differences between groups after controlling the level of the latent trait. In other words, systematic group differences only refer to differences in some statistical properties (e.g., mean scores) for a persons' membership in a group. However, measurement bias/systematic inaccuracy indicates differences in some statistical properties (e.g., mean scores) for members of different groups after controlling for the latent construct of interest.

Measurement invariance can also be conducted across time. Putnick and Bornstein (2016) pointed out that "Measurement invariance assesses the psychometric

equivalence of a construct across groups or measurement occasions and demonstrates that a construct has the same meaning to those groups or across repeated measurements." (p.72). Here, Putnick and Bornstein pointed out that testing measurement invariance longitudinally is also vital. In this situation, researchers are interested in studying individual changes in a latent variable over time. In order to ensure that the repeatedly measured variable has the same meaning over all time points, longitudinal measurement invariance need to be held. If the longitudinal measurement invariance does not hold, then, "the observed changes may reflect changes in what is being measured rather than in the level of the construct of interest" (Liu et al., 2017, p.486). Therefore, the evaluation of longitudinal measurement invariance is of importance in order to draw a valid conclusion about growth and change in the level of latent constructs across time points (Liu et al., 2017).

According to Raju, Laffitte, and Byrne (2002), "when measurement invariance is present, the relationship between the latent variable and the observed variable remains invariant across populations. In this case, the observed mean difference may be viewed as reflecting only the true difference between the populations" (p. 517). In other words, measurement invariance tests whether the equations used to create the latent factor scores are equal across groups (or across the continuous variables), ensuring that the constructs are operationalized similarly. Once an item, a set of items, or a latent construct is deemed invariant across groups, comparison of groups using parameter estimates (i.e., latent factor means or structural coefficients) is warranted. Generally, researchers could test "(a) the validity of the MI assumption, (b) the equality of latent factor means across

groups, and (c) whether the relationships between factors (structural coefficients) are equal across groups for a given theoretical model." (Sass, 2011).

2.1.2 Partial Invariance

Measurement invariance can be assessed across several levels of a measure, such as a single item, a set of items, a latent construct, or an entire measure. Ideally, researchers expect that full measurement invariance holds when all items' measurement parameters are the same across groups. In practice, however, this is rather difficult to achieve. When a subset of items in a measure is thought invariant, but the other small subset of items exhibits DIF, then partial invariance may be examined. Partial invariance is supported when at least two item parameters per construct are equal across populations (Cieciuch & Davidov, 2015). Although full invariance does not hold and group comparisons based on observed scores may yield misleading interpretation (Lai, et al., 2019), it is still sufficient and meaningful to make valid cross-group comparisons (i.e., means, intercepts, loadings) under partial invariance if item-level DIF is appropriately identified and the degree of non-invariance is small enough (Bauer, 2017; Bryne, et al. 1989; Lai, et al., 2019).

Recent simulation work conducted by Shi et al. (2019) has found that if all noninvariant parameters are correctly freed for estimation in the partially invariant model, the estimates of latent means, variances, inter-factor correlations, and coefficients of regressing factors on external variables are relatively accurate and consistent. More interestingly, Shi and colleagues (2019) further tested a single (correct) reference indicator (RI) model, where equality constraints were only placed on a one truly invariant item, and all other parameters including truly invariant items were freely estimated. The

result was intriguing, indicating that even if there is only one correct invariant item in a model, researchers can still obtain consistent estimations across latent means, variances, inter-factor correlations, and coefficients of regressing factors on external variables as long as that correct invariant item is identified correctly.

2.1.3 First-order Measurement Invariance

As stated above, measurement invariance concerns the conditional response distribution of the items. However, other weaker forms of invariance would also be useful in practice, such as the less stringent condition of first-order measurement invariance (Millsap, 2011, pp. 49-51). First-order measurement invariance is defined in terms of the conditional mean of the item responses as opposed to the conditional response distributions. Mathematically, this is defined as:

$$E(Y_i|T = t,\ V_i = v_i) = E(Y_i|T = t).$$

Under this form of invariance, the expected value is invariant across groups, but other statistical properties, such as the variance of response distribution may vary across groups. Comparing to the full measurement invariance, first-order measurement invariance is less stringent because the equation of $f_1(Y_i|T = t,\ V_i = v_i) = f_2(Y_i|T = t)$ may be violated, but would not necessarily be violated for the expectation of the same equation, that is: $E(Y_i|T = t,\ V_i = v_i) = E(Y_i|T = t)$. For example, if there is a great amount of variability in item responses of male and female students, then the expected values of math ability items may still be invariant across gender, but the response distribution for male students would differ from female students.

In general, measurement invariance is an important aspect of construct validity. It aims to ensure the measures produce a valid and comparable measurement (i.e.,

construct-level relationships) across all observations. Many theoretical and empirical research studies have been conducted on this topic within the context of the latent variable modeling framework. Scholars have established formal definitions of measurement invariance and depending on the presence of DIF or the satisfaction of full distribution of the item responses, different types of MI can be used in practical settings. The importance of assessing measurement invariance has increased rapidly since 1990 (Bauer, 2016). In addition, through the continued development of statistical software, methods of invariance testing have become more accessible to applied researchers. Given the greater prominence of measurement invariance in the field of measurement, a review and reexamination of the procedures used to assess MI is justified.

2.2 Testing Measurement Invariance Using the Multiple Groups CFA

In the social and behavioral sciences, multiple-group confirmatory factor analysis (CFA) is a popular analytic tool to address questions of validity during the process of instrument development. When using CFA models to evaluate cross-group equivalence, all measurement and structural parameters need to be examined across multiple groups. The measurement model aims to examine the equivalence of indicators including factor loadings, intercepts/thresholds, and residual variance. The structural model consists of the evaluation of the latent variables including factor means, variances, covariances, and regression coefficients (Brown, 2015). Generally, a common way to use CFA for testing multiple-group invariance is first to fit CFA models separately based on multiple populations or groups (e.g., gender, ethnicity). If the model fit is adequate within groups, then, a researcher may feel comfortable proceeding to test similarities and differences of factor structures and parameter estimates (factor loadings, intercepts/thresholds, variance,

and covariance). This is done by imposing equality constraints on model parameters. For example, one can set equality constraints for factor loadings across groups to identify optimal noninvariant parameter values that provide the best fit to the data. Finally, to decide on levels of invariance (e.g., configural, metric, or scalar invariance) across groups, one can compare model fit statistics of a series of nested multiple group models. When two models are nested, it indicates the more restricted model includes more restrictions or fewer free parameters than the less restricted model (Savalei et al., 2021).

In the step of comparison of nested models, researchers may use a Likelihood Ratio Test, also known as the chi-square difference test (Kline, 2011), or employ model fit indices to evaluate a change between a baseline model and a more restrictive model. While computing a chi-square difference test to compare nested models is popular in use in many applications, methodologists have found that when assessing differences in model fit, chi-square is too sensitive to sample size, and rejection of the null hypothesis is inaccurate when sample size is large (Chen, 2007; Savalei et al., 2021). In a recent review, Crede and Harms (2019) found that 90% of 112 reviewed articles conducting model comparisons used chi-square values but ignored the significance of chi-square values for the baseline model. Another problem of the chi-square difference test is that this approach is known as an exact fit, which aims to find a perfect fit or no discrepancies between the tested model and the model reproduced by the data (DiStefano, 2016), it may not be plausible to obtain exact fit in applied situations.

Alternatively, researchers may assess invariance of nested models using difference of model fit indices. As an increasing number of studies have investigated criteria on fit differences for evaluation of measurement invariance, this approach has

gained popularity in recent years (Crede & Harms, 2019). Based on literature, the most popular fit indices used for invariance evaluation are changes in RMSEA, and CFI (Cheung & Rensvold, 2002; Chen, 2007). However, most investigations are under the continuous data situation, studies related to fit indices in the context of invariance testing with Likert-type ordinal data are still rare.

CFA was originally developed for use with continuous outcome data, where model fit and parameters were usually estimated with the maximum likelihood (ML) estimation method. The ML estimation method requires the assumptions of independence and multivariate normality to produce accurate parameter estimates and standard errors (e.g., Bollen, 1989; Maydeu-Olivares, 2017). However, observed variables obtained in the social and behavioral sciences are often ordinal (i.e., Likert-type scale items) rather than continuous. As a result, significant problems may occur when ordinal scales are analyzed using ML estimation if the multivariate normality assumption does not hold (i.e., Muthén & Kaplan, 1985). While ML may be used in some situations when ordinal data are present, standard errors are no longer consistent if the assumption of multivariate normality is severely violated (Lubke & Muthén, 2004; Maydeu-Olivares, 2017). Therefore, alternative distribution-free estimation methods such as WLS, WLSM, or WLMV are needed to deal with this issue.

In the next section, I described the traditional case when continuous variables are used with MG-CFA, then, the use of MG-CFA invariance models with ordered data (e.g., Likert-type or discrete data) was introduced. Following these sections, another important concept, invariance constraints, was discussed. Lastly, I ended by addressing common estimation methods including ML, WLSMV, and fit evaluation of multiple-group CFA.

2.2.1 Measurement Invariance in Multiple-group CFA using Continuous Variables

The CFA model is a linear regression model where a large number of observed items are regressed on a small number of factors. When applied to multiple populations, the scores on the j^{th} continuously observed variable y_j in the k^{th} population can be denoted as follows:

$$y_j = \tau_{jk} + \lambda_{jk}\xi_k + \varepsilon_{jk}$$

Where τ_{jk} is an intercept (item mean) parameter for the j^{th} measured variable in the k^{th} population. λ_{jk} is the regression slope for the j^{th} measured variable in the k^{th} population and represents factor loading. ξ_k is the latent factor score and ε_{jk} is the unique factor score or residual for the j^{th} measured variable.

Recall that measurement invariance can be interpreted as invariance of the conditional distribution of observed scores (y_j) given the latent score (ξ_k) across groups. Thus, in the continuous CFA case, analysis of invariance should be based on the mean and covariance structures rather than correlation matrices. In addition, we also assume the following equations:

$$E_k(\varepsilon, k) = 0,\ Cov_k(\varepsilon, k) = \Theta_k$$

Where Θ_k is p x p covariance matrix of the residual scores of the observed items for the kth population, Θ_k is assumed to be diagonal, including only the item residual variance parameters $\sigma^2_{(11)k}$, …, $\sigma^2_{(pp)k}$. The covariance between the latent score and error is zero ($i.e., Cov_k(\xi, \varepsilon) = 0$) as is the covariance between error terms (i.e., $Cov(\varepsilon\varepsilon)$).

Using the assumption of multivariate normal distribution (MVN) underlying the observed scores (y_j), we can write the multiple groups CFA analysis model in terms of two sets of equations. The first set of equations expresses the conditional mean and

covariance structure given the latent variables and group membership. The second set of equations expresses the expected values and covariance structure of the latent construct in each group. These equations can be written as:

$$E_k(y_j|\xi, k) = \tau_k + \Lambda_k \xi,\ \ Cov_k(y_j|\xi, k) = \ \Theta_k$$

$$E_k(\xi, k) = \kappa_k,\ Cov_k(\xi, k) = \Phi_k$$

Where τ_k is a vector of intercepts, Λ_k is the group-specific p x r matrix of factor loadings, Θ_k is a variance-covariance matrix among the residuals, and κ_k is the r x 1 vector means for the latent factors. Φ_k is the covariance matrix of factor scores for the kth population. If measurement invariance across groups is satisfied, the conditional means and variance/covariance of observed scores given factor scores are independent of the group membership. Studies need to focus on investigating the invariance of parameters τ_k, Λ_k, and Θ_k across populations.

In addition to estimating the conditional mean and covariance structure of the observed variables (y_j) given the latent variables (ξ) in the kth population, unconditional equations for the observed variables (y_j) can be expressed as:

$$E_k(y_j, k) = \mu_k = \tau_k + \Lambda_k \kappa_k,$$

$$Cov_k(y_j, k) = \Sigma_k = \ \Lambda_k \Phi_k \Lambda_k' + \Theta_k$$

Where $E_k(y_j, k) = \mu_k$ is a vector of population means of the observed variables and Σ_k is the population variance-covariance matrix for the observed variables.

2.2.2 Measurement Invariance in Multiple-Group Testing with Ordered-categorical data

Thresholds of Ordered-categorical Data. The procedures used in the investigation of invariance in ordered-categorical measures is similar to the continuous data. For example, restrictive tests are employed to investigate different levels of

invariance. However, a major difference is that the threshold structure is embedded in the ordered-categorical CFA framework.

If we assume the ordered-categorical variable (y_j) is discrete, and there exists an underlying latent response variable (y_j^*) that is continuous and satisfies the assumption of underlying MVN (see Figure 2.1), then, the latent response variable (y_j^*) can be divided by a set number of categories (C), which is called threshold parameters (τ_j). The total number of thresholds is equal to the number of categories minus one (C-1) (Finney & DiStefano, 2013), and ordered-categorical responses (y_j) can be assigned values of 0,1, 2, …, C across all the populations. When C >1, the association between the underlying continuous latent variables at the observed level is denoted a polychoric correlation; when C = 1, it is denoted a tetrachoric correlation (a special case in which both ordinal variables have two categories) (Xia & Yang, 2019). Besides, the C-1 number of threshold values ($\tau_{jk} = \tau_{jk0}, \tau_{jk1}, \dots, \tau_{jkC-1}$) is within the range of ($\tau_{j0} = -\infty, \tau_{jC-1} = +\infty$). To clearly express the relationships among y_j, y_j^* and τ_{jk}, we can have the following equation when the observed variable is measured in multiple populations (Millsap, 2011):

$$P_k\ (y_j = C) = P_k\ (\tau_{jkC} \leq y_j^* \leq \tau_{jkC+1})$$

Let us use the example as presented in Finney and DiStefano (2013) to illustrate the relationships among y_j, y_j^* and τ_{jk}. Suppose that we have a five-point Likert scale item; thus, the observed level data (y_j) can only be reported as values from 1, 2, … to 5 (Figure 2.1). However, there exists an underlying continuous latent level response variable (y_j^*) that better represent the observed level data, (y_j). The relationship between y_j and y_j^* is connected by four (C-1) threshold values (τ_{jk}). As a result, instead of using

observed level data, our focus is to compute thresholds to obtain latent level variables. One important aspect is that the latent level response variable is assumed to be continuous and meet assumption of MVN; Additionally, the corresponding observed level data can be symmetric or asymmetric (Kim, 2012; Rhemtulla et al., 2012).

Non-normal Latent Response Distribution. Recent studies have extended their focus on investigating non-normal continuous latent level response distribution, in which the MVN assumption is violated (e.g., DiStefano, 2002; Liang, & Yang, 2014; Muthén & Muthén, 2002; Pavlov et al., 2020). For example, Flora and Curran (2004) examined the normal distribution and moderate non-normal distribution of latent responses with skewness of up to 1.25 and kurtosis of up to 3.75. Using both full WLS estimator and robust WLS estimator, they found that increasing levels of non-normality in latent response distributions were related to a greater positive bias in estimated polychoric correlations and parameter estimates. However, the level of bias remained low for the moderate non-normal latent response distribution.

Rhemtulla et al. (2012) extended Flora and Curran's study to examine the effect of nonnormality levels of continuous latent response distribution and threshold variability (i.e., symmetric or asymmetric) on the performance of ML and categorical least squares (cat-LS) estimation. They concluded that compared to the robust ML estimation, cat-LS estimation is more sensitive to the violations of non-normality of underlying continuous variables and was superior to ML with two to four categories with mild bias in underlying non-normal distribution, asymmetric thresholds, and small sample sizes.

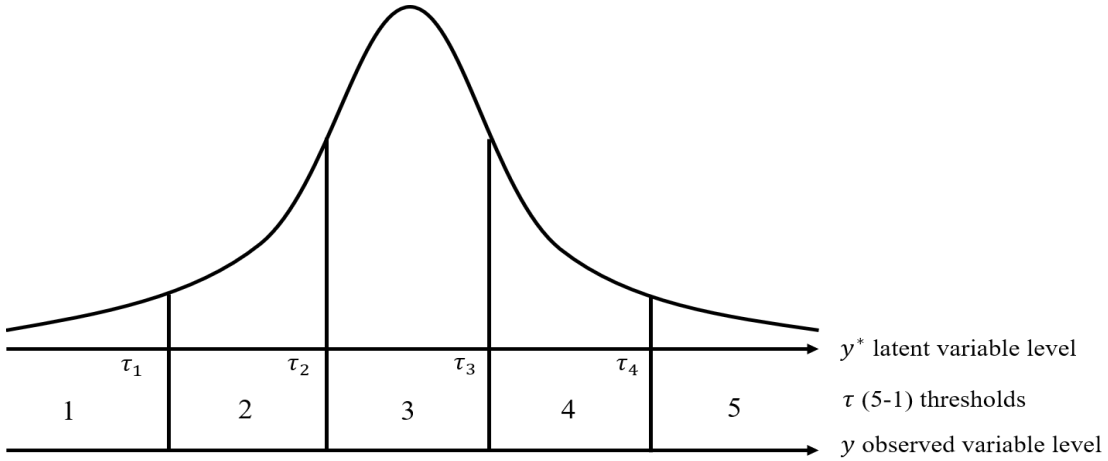

Figure 2.1 Relation between y_j, y_j^* and thresholds. Adopted from (Finney and DiStefano, 2013).

Multiple-Group CFA with Ordered-categorical data. As shown in Figure 2.2, the confirmatory factor analysis model is no longer directly related to the observed response variables (y_j) when ordinal data are used but associates with y through the underlying latent response variables (y_j^*). According to Kim and Yoon (2011), the expression for the conditional mean and covariance structure for the CFA model with continuous latent response variables y_j^* and discrete observed scores y_j in the k^{th} population can be written as:

$$E_k(y_j^*|\xi, k) = \tau_k + \Lambda_k\,\xi,\ \ Cov_k(y_j^*|\xi, k) = \ \Theta_k$$

Where ξ is the r x 1 vector of latent factor scores, Θ_k is p x p diagonal matrix of residual variance for the kth population, τ_k is the p x 1 vector of latent intercept parameters for the kth population, and Λ_k is the p x r factor loadings matrix.

Assuming MVN of the latent level responses (y*), the score-level expression for the observed variables (y_j) in the k^{th} population can be expressed as:

$$y_j^* = \tau_{jk} + \lambda_{jk}\xi_k + \varepsilon_{jk}$$

Where ε_{jk} is a p x 1 vector of the residual score, and we still assume that $E_k(\varepsilon, k) = 0$ and $Cov_k(\varepsilon, k) = \Theta_k$. For the common factor ξ_k , we still have the following equation that is identical to the continuous multiple-population CFA models:

$$E_k(\xi, k) = \kappa_k,\ Cov_k(\xi, k) = \Phi_k;$$

$$E_k(y_j^*, k) = \mu_k^* = \tau_k + \Lambda_k \kappa_k;$$

$$Cov_k(y_j^*, k) = \Sigma_k^* = \Lambda_k \Phi_k \Lambda_k' + \Theta_k$$

As the thresholds determine the distribution of responses on ordered-categorical variables, testing measurement invariance with multiple-group CFA for ordinal measures aims to estimate the invariance of thresholds $(\tau_{jk0}, \tau_{jk1}, \dots, \tau_{jkC-1})$, and parameters of Λ_k, Θ_k , as with in MG-CFA with continuous data.

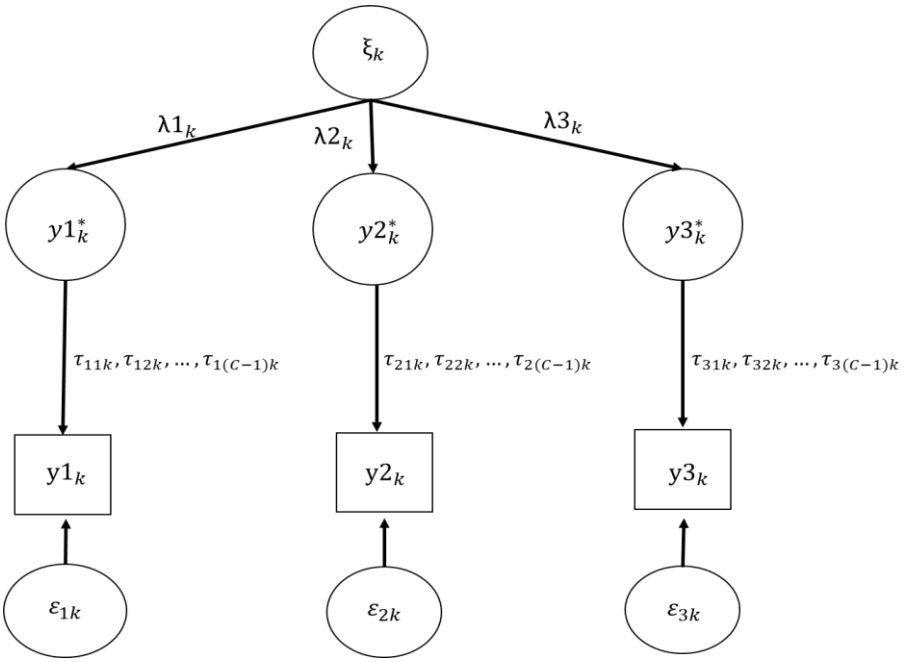

Figure 2.2 Multiple-group categorical CFA models.

2.2.3 Invariance Constraints

To address whether any differences exist in the mean and covariance structures of the observed variables across populations, Meredith (1993) defined a hierarchical set of

invariance comparisons in the context of the CFA approach. According to Meredith (1993), the lowest level of invariance is configural invariance, which requires that the factor structure holds the same in each group, but no invariance constraints are placed on factor loadings, item intercepts/thresholds, and item residual variances. The model tested is:

$$H_{01}= \Sigma_k/ \Sigma_k^* = \Lambda_{kc}\, \Phi_k\, \Lambda'_{kc} + \Theta_k,\ \ \mu_k/ \mu_k^* = \tau_k + \Lambda_{kc}\kappa_k$$

for k = 1, …, K^{th} group, where Σ_k/Σ_k^* is the population covariance matrices for either continuous or ordered observed variables, μ_k/μ_k^* is the population mean vectors of observed variables, and Λ_{kc} refers to the factor loading matrices. This indicates that the loading matrices have the independent cluster structure, c, across the K groups. The model in H_{01} indicates that the number of factors is the same across groups/populations and the factors are related to the same number of items in each group/population. Rejection of H_{01} indicates that the factor structure is untenable for at least one or more groups/populations. For additional information, a technical discussion can be found in Meredith (1993).

If H_{01} cannot be rejected, this implies that the same factor structure holds, the next level of invariance tested is weak factorial invariance or metric invariance (Horn & McArdle, 1992; Thurstone, 1947). In this step, only the factor loadings are constrained to be equal across groups ($\Lambda_k = \Lambda$), the intercepts/thresholds and residual variances are free to vary:

$$H_{02}= \Sigma_k/ \Sigma_k^* = \Lambda\, \Phi_k\, \Lambda' + \Theta_k,\ \ \mu_k/ \mu_k^* = \tau_k + \Lambda\, \kappa_k$$

for k = 1, …, K. This implies that the factor loadings are the same over groups. In this level of invariance, the mean of the latent factor is fixed at zero in group one and

estimated in the other groups. Note that weak invariance satisfies neither the full invariance of the conditional distribution function nor the first order invariance. It does, however, permit comparisons of factor variances and covariances.

Assuming that H_{02} cannot be rejected, the next question of interest would be to decide whether a more stringent invariance can be obtained. The next level of invariance then is called strong factorial invariance or scalar invariance, which requires that intercepts/thresholds and factor loadings are equal ($\tau_k = \tau;\ \Lambda_k = \Lambda$). The residual variances of the items, however, are free to differ (Steenkamp & Baumgartner, 1998):

$$H_{03} = \Sigma_k / \Sigma_k^* = \Lambda\, \Phi_k \Lambda' + \Theta_k,\ \ \mu_k / \mu_k^* = \tau + \Lambda\, \kappa_k$$

for k = 1, …, K. The model in H_{03} meets the first-order invariance condition, and places testable restrictions on the means of the observed variables. If H_{03} holds, the group membership has no impact on the expected values of the items conditioning on the latent variables. Under this condition, comparisons of factor means are possible across groups. If H_{03} does not hold, the factor loadings and/or intercepts (or thresholds) contribute differently to the means. As a result, it prevents valid and comparable factor score estimates (Sass, 2011). Strong invariance is often considered sufficient in most empirical research (Bauer, 2017).

The highest level of invariance is strict factorial invariance, which requires that all item parameters ($\tau_k = \tau;\ \Lambda_k = \Lambda;\ \Theta_k = \Theta$) are equal across groups. Specifically, the model being tested is:

$$H_{04} = \Sigma_k / \Sigma_k^* = \Lambda\, \Phi_k\, \Lambda' + \Theta,\ \ \mu_k / \mu_k^* = \tau + \Lambda\, \kappa_k$$

for k = 1, …, K. In this way, the response distributions are independent of group membership after conditioning on the values of the latent factors. Under strict invariance,

group differences in the means and covariance structure for the observed variables are due to differences in latent variable distributions, rather than measurement bias (Millsap, 2011). Failure to reject H_{04} indicates no measurement bias is detected. Although strict factorial invariance is considered a necessary condition of measurement invariance, it is often too restrictive to be met in most empirical research situations.

Measurement invariance can be tested at more advanced levels as well, such as factor variances and covariances, or factor means can be constrained across groups. However, the stricter levels of invariance are not often accomplished in practice. Turns out, configural, weak, and strong factorial invariance are the most commonly tested forms of invariance in applied research for both continuous and ordinal data (Widaman & Reise, 1997).

2.2.4 Estimation Methods for MI

In terms of SEM model estimation, when data are continuous, the Maximum Likelihood (ML) estimation is generally utilized. ML estimation is based on a large sample size and multivariate normality assumptions because ML depends on satisfying the distribution assumptions for observed variables to obtain adequate performance. To obtain estimated parameters ($\tau_k, \Lambda_k, \Theta_k, \kappa_k, \Phi_k$) of CFA models, the goal is to minimize the differences between the true and observed scores, which is to minimize the discrepancy functions F_{ML} ($\overline{X}_k, S_{Xk,} \mu_{Xk,} \Sigma_{Xk,}$), where $\overline{X}_k$ is the sample estimator of μ_{yk}, and S_{yk} is the sample estimator of Σ_{yk} . The form of the discrepancy function varies depending on the method. Under the MVN assumption, the normal maximum likelihood discrepancy function for the multiple-group case is (Millsap, 2011):

$$F_{ML} = \sum_{k=1}^{k} \left(\frac{N_k}{N}\right) F_{MLk}$$

Where F_{MLk} is defined as:

$$F_{MLk} = \left(\bar{X}_k - \mu_{yk}\right)' \sum_{yk}^{-1} \left(\bar{X}_k - \mu_{yk}\right) + \ln \frac{\left|\Sigma_{yk}\right|}{\left|S_{yk}\right|} + \mathrm{tr}\left[\sum_{yk}^{-1} S_{yk}\right] - p$$

The discrepancy function ranges from zero to infinity, and larger values of F_{ML} indicate greater discrepancies between the observed and implied covariance matrices. When invariance constraints are imposed, software packages such as LISREL, EQS, R, or Mplus are used to estimate multiple-group analyses. The discrepancy function value at the minimum of F_{ML} is used for calculating a chi-square test for the null hypothesis that a specified model fits in the K groups (Millsap, 2011). Specifically, the chi-square test statistic is formed by:$\chi^2 = F_{ML} * (N - 1)$, and this test statistic follows a central chi-square distribution. When the null hypothesis fails, the discrepancy function can help to calculate the noncentrality parameter, which is an important component of a noncentral chi-square distribution and carries important information about the degree of model misspecification (Curran, 2002). The noncentral chi-square distribution helps to construct fit indices such as RMSEA, CFI, or TLI.

When the data are ordered categorical, the threshold parameters (τ_k), means (μ_k^*) and covariance matrices (Σ_k^*) of the latent response variable (y_j^*) need to be estimated. Usually, we assume that the latent response variable y_{jk}^* follows:

$$y_{jk}^* \sim \mathrm{MVN}\left(\mu_k^*, \Sigma_k^*\right)$$

Holding the MVN assumption and appropriate restrictions, estimates of (τ_k, μ_k^*, Σ_k^*) can still be obtained by maximum likelihood.

However, to provide estimates of the factor model parameters ($\Lambda_k, \kappa_k, \Phi_k, \Theta_k$) in multiple groups under invariance constraints, many other estimation approaches (e.g., robust maximum likelihood; MLR, weighted least squares; WLS, or diagonally weighted least squares; DWLS) have been developed to consider the non-normality nature of ordinal data. While there are many choices, I focus on reviewing research that employs the weighted least squares mean and variance adjusted (WLSMV) estimator. This is the most popular estimator for ordered categorical data and has been recommended for estimating CFA model parameters when data is ordered categorical (Muthén & Muthén, 2010). Researchers, who are interested in other estimation methods used with non-normal data can find details in many other articles such as Yuan and Bentler (2000), Finney and DiStefano (2013), Millsap (2011), Flora and Curran (2004).

Weighted Least Squares Mean and Variance Adjusted (WLSMV) Estimator. WLSMV estimation, which is called weighted least squares mean and variance adjusted estimator, was originally from robust Diagonally WLS (DWLS) estimation and was first introduced by Muthén in 1993. The WLSMV estimator is designed specifically for noncontinuous data that multivariate normality assumption may be violated (DiStefano & Morgan, 2014). To analyze ordinal observed data, the WLSMV estimator first estimates thresholds and polychoric correlation (correlation between two ordinal variables) using ML estimation, and the parameter estimates are then calculated by minimizing the discrepancy function F_{WLSMV} using the estimated asymptotic covariance matrix of the polychoric correlation, as well as the threshold estimates in a diagonal weight matrix (Li, 2016). The multiple-group WLSMV discrepancy function can be written as:

$$F_{WLSMV} = \sum_{k=1}^{k} \left(\frac{N_k}{N}\right) F_{WLSMVk} ,$$

and the equation to minimize WLSMV function is:

$$F_{WLSMVk} = (r_k - \hat{\rho}_k)(diagW^{-1})(r_k - \hat{\rho}_k)$$

Where r_k represents a vector of unique elements in the sample covariance matrix (S) for the kth group including threshold and polychoric correlation estimates, $\hat{\rho}_k$ represents a vector of the nonduplicated elements in the model-implied covariance matrix $[\sum(\hat{\theta})]$ for the k^{th} group, and $r_k - \hat{\rho}_k$ is a residual vector of the discrepancies between the sample values and model-implied values. diag**W** is the diagonal weight matrix, which utilized the asymptotic covariance matrix of the polychoric correlation estimates and thresholds (Finny & DiStefano, 2013).

The diag**W** is a special weight matrix form used for the WLSMV estimator, and it adjusts the departure from normality and sampling variably in the formula. Since there are several practical problems in implementing WLS estimation (see Finney & DiStefano, 2013), WLSMV is developed to overcome the limitations of full WLS estimation. According to DiStefano and Morgan (2014), unlike the WLS estimator which inverts the full weight matrix, WLSMV only inverts the diagonal elements of the weight matrix, so the computational intensity is decreased to avoid using a large sample size.

A mean- and variance-adjusted chi-square test statistic with the degrees of freedom is computed based on the formula (Muthén, 1993; Li, 2016):

$$T_{DWLSMV} = [df'/ \text{trace } (\tilde{U}\tilde{V})]\, T_{DWLS},$$

Where $\tilde{V}$ is the estimated asymptotic covariance matrix of the thresholds and polychoric correlations, $\tilde{U} = diagW^{-1} - diagW^{-1}\tilde{\Delta}\left(\tilde{\Delta}'diagW^{-1}\tilde{\Delta}\right)^{-1}\tilde{\Delta}'diagW^{-1}$, and $\tilde{\Delta} = \partial\sigma(\tilde{\theta})/\partial\tilde{\theta}$, $\tilde{\theta}$ is a vector of the estimated model parameters, df' is an integer closet to $\{[\text{trace } (\tilde{U}\tilde{V})]^2$

/trace $(\tilde{U}\tilde{V})^2\}$[1], T_{DWLS} is the standard (N-1) the minimum of the fit function (see details in Li, 2016). By adjusting mean and variance for test statistics using the WLSMV method, one can obtain an approximate chi-square distribution with associate degrees of freedom, which is used for fit evaluation.

In general, many studies have been found that WLSMV provided fairly accurate parameter estimates (e.g., Flora & Curran, 2004; Bandalos & Webb, 2005) and works effectively in most situations when ordered categorical variables are used with CFA for sample sizes larger than 500 (Bandalos, 2008; Flora & Curran, 2004; Muthén & Muthén, 2010). Comparing to the other estimators, studies have shown that with simulated data, WLSMV provided a less biased and more accurate estimation of factor loadings across almost every condition than robust ML (MLR) (Li, 2016). Flora and Curran (2004) found that WLSMV outperformed the WLS estimator in chi-square approximation of the test and having smaller estimation biases of the parameters when using complex CFA models. DiStefano and Morgan (2014) concluded that WLSMV produced better model data fit than WLSM. They also found that WLSMV (Mplus version) was a better choice than DWLS (LISREL version) with small sample sizes, few categories, and moderate sample distribution (skewness=1.5, and kurtosis =3). Overall, using the WLSMV estimator with ordinal data has been found to be superior over many estimators and has been found to produce adequate parameter estimates in many simulated conditions (DiStefano & Morgan, 2014; Schmitt, 2011).

[1] See Li (2016) for details of formulas and expressions.

2.2.5 Fit Evaluation for Measurement Invariance

When conducting a multiple-group CFA invariance test, an important step is to examine the model fit for each group to make sure that the CFA model fits acceptably across groups (Sass, 2011). In general, an adequate CFA model for each group should be satisfied the first step. Then, a series of tests (i.e., likelihood ratio tests) can be conducted to compare parameters (e.g., factor loadings, intercepts) of a baseline model with more restrictive nested models based on different levels of invariance constraints. After that, a series of model fit statistics (i.e., chi-square statistics) need to be evaluated to determine the magnitude of model differences. For example, when a researcher examines weak invariance, one needs to constrain all factor loadings to be equal across groups and evaluate the significance of the chi-square difference between this model and baseline model. If the chi-square difference test result indicates non-significance, then the model with more restricted constraints performs as well as the baseline model. Thus, further constraints can be added to test a higher level of invariance (i.e., strong invariance). If the chi-square difference test is significant, researchers may set free constraints of non-invariant factor loadings and carry out the partial invariance method.

As stated, although the chi-square difference test is commonly used, it has several drawbacks. Instead of using the chi-square difference test, many fit indices such as RMSEA, CFI, SRMR have been proposed to evaluate measurement invariance. Unfortunately, studies of fit indices for invariance evaluation with ordered categorical data are unclear. This study aimed to fill the gap in the literature by examining commonly used model fit indices when to evaluate measurement invariance with ordered categorical data. I first reviewed several goodnesses of fit indices that are most used with multiple-

group CFA invariance testing including both traditional and scaled χ^2 tests, and other fit indices, such as RMSEA, CFI, and SRMR. Then, a limited number of articles that examined the sensitivity of goodness of fit indices to lack of measurement invariance for both continuous and ordered-categorical measures were discussed (e.g., Cheun & Rensvold, 2002; Chen, 2007; Rutkowski & Svetina, 2014; Rutkowski & Svetina, 2017).

Chi-square (χ2). This exact fit test measures the entire model-data fit using statistical hypothesis tests across multiple populations. The chi-square goodness-of-fit is widely used for continuous CFA models with the maximum likelihood estimator, and its calculation is based on the discrepancy between the actual model's covariance matrix (Σ_y), mean structure (μ_{yk}), the hypothesized model's covariance (Σ_{0yk}) and mean (μ_{0yk}) structure in the k^{th} group. The null hypothesis H_0: $\Sigma_{yk}=\Sigma_{0yk}$, $\mu_{yk}= \mu_{0yk}$, for k=1,2, …, k will be tested. Then, the chi-square statistic is calculated: $\chi^2= (N-1)\ \hat{F}$, where $\hat{F}$ is the minimized sample discrepancy function value, N is the sample size, and the degrees of freedom are the differences between the number of residual variances, covariance, mean elements, and independent parameters (Millsap, 2011). Specifically, when testing measurement invariance with two models, A and B, a difference of the chi-square statistic between the two CFA models is calculated, by finding the difference between values for model B and model A, where model B is nested within model A. That is, model B is the more restrictive model with more degrees of freedom than the comparison model. Therefore, we can write the equation as:

$$\chi^2_D=\chi^2_A-\chi^2_B \text{ with } df_D = df_B - df_A$$

Assuming both models fit well, then the difference in the chi-square can be used to detect whether model B is a lack of fit in comparison to model A. The significant

difference test can then examine whether or not the difference between model B and model A is statistically significant. As the chi-square difference test is obtained from $\hat{F}$, which usually indicates the maximum likelihood discrepancy function value, the assumptions of multivariate normality are required when using this test statistic. When data are not normally distributed, the difference of fit statistics of two nested models does not result in a chi-square distribution. In such case, the Satorra-Bentler (SB) scaled chi-square difference test (Bryant & Satorra, 2012; Satorra & Bentler, 2001, 2010) or more advanced chi-square correction difference tests for categorical non-normal data (i.e., Asparouhov & Muthén, 2006, 2010) may be used.

Second-order Chi-square (χ2) Correction. Asparouhov and Muthén (2010) proposed a new second-order correction statistic, the T3 method, to deal with multivariate nonnormality and to transform the fit difference between two nested models more similar to a chi-square distribution. This technique has been implemented in Mplus Version 6 with estimators WLSMV and MLMV under the "DIFFTEST" command. The second-order correction is designed to match both the mean and variance of the chi-square distribution with D degrees of freedom, where D is the difference between the number of parameters in the unrestrictive model and the estimated model. The form of the mean and variance adjustment takes $T_3 = aT + b$, where a is a scaling correction, T is the chi-square difference between two models, and b is a shifting parameter. Both a and b are chosen to meet E (T_3) = D, Var (T_3) = 2D. The second-order correction (Asparouhov & Muthén, 2010) is given by

$$T_3 = \sqrt{\frac{D}{Tr(M^2)}}T + D - \sqrt{\frac{DTr(M^2)}{Tr(M^2)}}$$

Where M is a matrix and is given in formula (9) in Asparouhov & Muthén (2006)[2]. The mean and variance of T_3 are the same with the chi-square distribution with D degrees of freedom.

To apply T_3 second-order correction for chi-square difference testing, we can use the "DIFFTEST" command in Mplus. Suppose two nested models A and B, where A is the more restricted model and B is the less restrictive model. The difference in fit between two nested models can be tested by subtracting the two fit statistics using T_3:

$$T_d=T_A-T_B$$

Since the distribution of T_d is not a chi-square distribution, which we cannot use the P-value directly. To achieve T_3 second-order correction difference test results, we can use T_d to approximate a chi-square distribution with D degrees of freedom.

According to Asparouhov and Muthén (2010), the new second-order correction is more advantageous to the old version second-order correction (Satterhwaite, 1941) because the degrees of freedom are not needed to estimate, and researchers can simply use the difference between the number of parameters in the two models to replace the usual degrees of freedom. For more technical information on the calculation of the new second-order correction for chi-square difference testing, readers can review Asparouhov and Muthén (2010).

Overall, the global chi-square statistic is classified into exact fit indices that assess the degree to which the model-implied covariance matrix matches the observed covariance matrix. It provides one single index to summarize the fit of the entire model and gauge the discrepancy or "badness of fit", therefore the smaller the number is, the

[2] See Asparouhov & Muthén (2006) for details of formulas and expressions.

better the model fit. Traditional chi-square statistic is useful when continuous data are employed in the research and when the normality assumption is met. The scaled chi-square statistic and the new second-order correction difference test are more robust than the traditional chi-square approach for non-normal continuous data or ordered-categorical data. However, a difficulty is that all the chi-square tests are sensitive to sample size. As it turns out, rejection of the null hypothesis of the exact fit can be easily obtained when the sample size is large (e.g., Chen, 2007). Additionally, in most of the studies, we often are less interested in the models that fail to fit perfectly than the extent and location of the misfit (Millsap, 2011). Therefore, other model fit indices such as CFI, RMSEA, or SRMR can help us quantify the size of the misfit.

Root mean squared error of approximation (RMSEA). The second type of absolute fit index is the root mean square error of approximation (RMSEA) (Steiger & Lind, 1980; Browne & Cudeck, 1993). This approach measures how far the hypothesized model is from the perfect fit to the data (McDonald & Ho, 2002). It considers both covariance and mean structures when the discrepancy function includes mean structures in the estimation. The RMSEA uses the information obtained from the discrepancy function $\hat{F}$ to estimate a closeness between $(\Sigma_{0Xk}, \mu_{0Xk})$ and (Σ_{Xk}, μ_{Xk}). Mathematically, the RMSEA is defined in one single population as:

$$\mathrm{RMSEA}_1 = \sqrt{\frac{\tilde{F}}{df}}$$

Where $\tilde{F}$ is the minimized fit function of the hypothesized model at the population level, df is the model's degrees of freedom. When WLSMV is used, the mean-and variance-

adjusted chi-square is applied to compute RMSEA fit in software packages such as Mplus. Therefore, the scaled RMSEA is calculated as

$$\mathrm{RMSEA_{ss}} = \sqrt{\frac{\hat{a}(N-1)\hat{F} + \hat{b}}{df(N-1)} - \frac{1}{N-1}}$$

Where $\hat{a}$ and $\hat{b}$ converge to a and b (the scaling parameter and shifting parameter; Asparouhov & Muthén, 2010); and $\hat{F}$ converges to $\tilde{F}$ as N increases to infinity, which converges to (Xia & Yang, 2019)

$$\mathrm{RMSEA_s} = \sqrt{\frac{a\tilde{F}}{df}}$$

Note that the RMSEA evaluates $\tilde{F}$ relative to the degrees of freedom, which penalizes models that include unnecessary parameters (Hu & Bentler, 1998). However, RMSEA tends to over-reject a true model when the sample size is small and is not recommended when evaluating small sample size models with small degrees of freedom (Kenny et al., 2015). Steiger (1998) extended the single population RMSEA to multiple groups (K) RMSEA by using a correction parameter. Thus, the modified formula is denoted as:

$$\mathrm{RMSEA_K} = \sqrt{K}\sqrt{\frac{\tilde{F}}{df}}$$

When using RMSEA as an overall fit index for evaluation of SEM models with a single group or population, Browne and Cudeck (1993) suggested values below 0.05 are considered as a good fit, and values between 0.05 and 0.08 indicate a fair fit. Hu and Bentler (1999) recommended that adequately fitting models should have RMSEA values below 0.06. MacCallum et al. (1996) used 0.01, 0.05, and 0.08 to indicate excellent,

good, and mediocre fit. These authors also suggested using confidence intervals as a supplement to the point estimate of the RMSEA. A 95% confidence interval for the single group RMSEA can be written as:

$$(\hat{\varepsilon}_{.025}, \hat{\varepsilon}_{.975}) = \left(\sqrt{\left[\frac{\hat{\lambda}_{.025}}{\mathrm{df}(N-1)}\right]}, \sqrt{\left[\frac{\hat{\lambda}_{.975}}{\mathrm{df}(N-1)}\right]}\right)$$

Where $\hat{\varepsilon}_c$ is the estimated bound for the RMSEA at the (100 * c) percentile, and $\hat{\lambda}_c$ is the estimated bound for the noncentrality parameter of the non-central χ^2 distribution with degrees of freedom (df) and N-1.

In terms of criteria used to evaluating measurement invariance with RMSEA, the above criteria are still applicable for configural invariance testing. However, investigations on changes in RMSEA for other invariance levels of nested models are insufficient. Only several researchers conducted simulation studies to propose criteria on changes of fit indices in cross-sectional CFA models in multiple-group cases (i.e., Chen, 2007; Cheun, & Rensvold, 2002; Rutkowski & Svetina, 2014; 2017). Chen (2007) recommended for all three levels of invariance tests (loadings, intercepts, residual variances): a change in the RMSEA greater than or equal to 0.01 (when sample sizes are unequal in groups and the sample size is smaller than 300) or 0.015 (when the sample size is adequate, sample sizes are qual across the groups) indicates measurement invariance is violated. The finding also found that changes in RMSEA are more likely to be affected by sample size and model complexity.

Rutkowski and Svetina's (2014) recommendation is more liberal when the number of groups was relatively large, and the indicators were assumed to follow a multivariate normal distribution. Specifically, the authors recommended that a change of RMSEA is no larger than 0.03 for tests of metric invariance or equal loadings, and the

traditional cutoff of 0.01 still works well at identifying scalar invariance. Rutkowski and Svetina (2017) also provided criteria of non-invariance with categorical indicators for large numbers of groups and non-normal observed variables: a change in the RMSEA equal to or greater than 0.05 for testing loading non-invariance/ metric non-invariance, and 0.01 for threshold non-invariance/ scalar non-invariance.

Comparative fit index (CFI). The third approach of global fit evaluation is the comparative fit index (CFI) that assesses the fit of the specified model relative to a more restricted baseline model. CFI fit index is classified as the incremental fit indices, which assess the degree to which the tested model is superior to a baseline model. Therefore, the larger the fit number is, the better the model fit. Larger values mean greater improvement of model fit comparing to the other model. Usually, the baseline model Σ_{yk} is equal to a diagonal matrix, D_{yk}. That is, the measured variables are mutually uncorrelated. In other words, a typical baseline model is the one in which only the variances of the observed variables are estimated, but no covariances are calculated. The Comparative Fit Index (CFI) is calculated as (Bentler, 1990):

$$\text{CFI}= 1 - \frac{\tilde{F}_t}{\tilde{F}_b} = 1 - \left\{\frac{{\chi_t}^2 - df_t}{{\chi_b}^2 - df_b}\right\}$$

Where $\tilde{F}_t$ and $\tilde{F}_b$ are the minimized fit functions of the tested and baseline models, ${\chi_t}^2$ is the chi-square for the tested model, ${\chi_b}^2$ is the chi-square for the baseline model, and df_t, df_b are the degrees of freedom for the tested model and baseline model, respectively. CFI ranges from 0 to 1, and it is relatively independent of sample size and performs well when the sample size is small (Hu & Bentler, 1998). According to Hu and Bentler (1995), values of the CFI above 0.95 were considered to be a good fit. Like the scaled RMSEA, the scaled CFI at the sample level is calculated with WLSMV as

$$CFI_{ss} = 1 - \frac{\hat{a}_t(N-1)\hat{F}_t + \hat{b}_t - df_t}{\hat{a}_b(N-1)\hat{F}_b + \hat{b}_b - df_b}$$

When the sample size increases to infinity, the equation converges to

$$CFI_s = 1 - \frac{a_t\tilde{F}_t}{a_b\tilde{F}_b}$$

Although research interested in fit statistics is growing for the factorial model (e.g., Bandalos, 2008; Beauducel & Herzberg, 2006; Yu & Muthén, 2002), these studies have not resulted in the same attention being given to addressing criteria of CFI fit statistics used to evaluate measurement invariance. This is especially true when ordered-categorical indicators are used. In research examining the performance of the CFI fit index for measurement invariance, Cheung and Rensvold (2002) examined the sampling variation of changes in the CFI index under various levels of measurement invariance. They suggested that a change in CFI (≥ -0.01) is sufficient for establishing weak or strong invariance. Similarly, French and Finch (2006) investigated measurement invariance with first-order models. Their simulation results recommend a ΔCFI value less than -0.01, indicating a lack of invariance. Chen (2007) conducted an extensive simulation study to examine the sensitivity of goodness of fit indices to lack of measurement invariance with continuous multiple-group CFA models at three common levels: factor loadings, intercepts, and residual variances. She found that CFI appeared to be equally sensitive to all three levels of lack of invariance and recommended that when the sample size is small (N≤ 300), sample sizes are unequal across groups, and the pattern of non-invariance is uniform, the cutoff criterion for a change of CFI at three levels of invariance tests is a change of ≤ -0.005 to indicate non-invariance. If the sample size is large enough (>300),

sample sizes are equal across the groups, and the lack of invariance is mixed, a change of ≤ -0.010 in CFI is indicative of non-invariance.

Rutkowski and Svetina (2014) recommended a change of less than -0.02 in CFI for metric invariance when the group sizes are large. In a follow-up study, Rutkowski and Svetina (2017) examined measurement invariance with categorical indicators. They found that in all conditions with non-invariant loadings, the changes of CFI values were not less than -0.012. However, a more stringent change in CFI of -0.004 is recommended for testing equal loadings. According to the results, they found that if using -0.012 as a criterion, ΔCFI could not detect poor-fitting models, but retained both well-fitting and poor-fitting models. In terms of tests of equal loadings and thresholds, they also suggested using the criteria of -0.004 to detect invariance for the same reason.

Standardized root mean square residual (SRMR). Lastly, I reviewed studies employing the standardized root mean square residual (SRMR) fit statistic. The SRMR is an approximate fit index that is designed to compare a hypothesized model and a baseline model (i.e., a model assuming zero correlation between every pair of variables), and its definition varies across publications (i.e., Asparouhov & Muthén, 2018; Maydeu-Olivares, 2017; Hu & Bentler,1999). According to Asparouhov and Muthén (2018), the SRMR fit index has been improved to use for more SEM models including models with categorical data estimated using WLS/ WLSM/WLSMV/ ULSMV in Mplus since version 8.1. As a residual-based fit index, SRMR computes the average differences of the standardized residuals between the observed and model-implied covariance matrices. The smaller the residuals, the better the model fit. The formula used to calculate SRMR with WLSMV estimator for categorical data is defined as:

$$\mathrm{SRMR} = \sqrt{\frac{S}{d}}$$

Where the detailed calculation of S and d can be found in formulas (13)-(16) in Asparouhov and Muthén (2018). In multiple group modeling, the SRMR_g is computed for each group g =1, …, G where G is the total number of groups. Then, the SRMR for the full model is denoted as follows

$$\mathrm{SRMR} = \sum_{g=1}^{G} \frac{n_g}{n} \mathrm{SRMR}_g$$

Where n_g is the sample size for group g, and n is the total sample size calculated by $\sum_{g=1}^{G} \frac{n_g}{n}$. According to Asparouhov and Muthén (2018), the SRMR is not a test, but a value that measures the direct distance between the hypothesized model and the baseline model. Therefore, it is easy to interpret and can be applied to identify model misfits.

To evaluate an overall fit of the SEM model, the SRMR index and the chi-square test of fit are paired in use (Asparouhov & Muthén, 2018). At the first, researchers should look at the chi-square test of fit, and if the chi-square fit does not hold, then the SRMR index should be used. The acceptable range of a good fitting model for the SRMR index is between 0 and 0.08 (Hu & Bentler, 1999).

When evaluating measurement invariance, Chen (2007)'s simulation study recommended that different values of the SRMR should be used based on different levels of invariance tests because SRMR is more sensitive to non-invariance in loadings than intercepts or residual variances. Specifically, a change of ≥ 0.025 in SRMR was proposed to indicate non-invariance for testing loading invariance; a change of ≥ 0.05 in SRMR would indicate non-invariance for testing intercept or residual invariance. All these cutoff

criteria are suggested with an unequal sample size less or equal to 300 and the pattern of non-invariance is uniform. When the sample size is over 300 and equally across the groups with mixed non-invariance items, Chen (2007) suggested using a change of ≤ 0.03 in SRMR for testing loading invariance and a change of ≤ 0.01 for testing intercept or residual invariance.

2.3 Current Research Gaps

In summary, while substantial research has emphasized the importance of detecting measurement invariance to ensure the validity of a measure and many model fit indices have been discussed by researchers, I found that limited scholarly attention has been given to the examination of the performance of changes in fit statistics to measurement invariance or lack of measurement invariance. Further, to my knowledge, only one study has been investigated to examine the performance of change in SRMR fit index to measurement invariance, and the sensitivity of the change in SRMR fit index to a lack of measurement invariance with ordered-categorical data is unknown. Many questions about this topic need to be addressed, such as,

1. What standards should be used to assess invariance using various fit indexes with ordered categorical data?;
2. Can uniform standards be proposed for testing measurement invariance at all levels (thresholds, loadings, and residual variances) with ordered categorical data?;
3. Are the fit invariance criteria proposed for both continuous and ordered data consistent?; and

4. Are there any factors such as the number of factors, magnitudes of factor loadings, and sample size impacting the fit invariance criteria?

Currently, many applied researchers use likelihood ratio difference test to detect the non-invariance of nested models. However, as stated, the use of chi-square-based tests has been questioned. Studies are needed to evaluate the performance of other model fit indices when models are non-invariant. Although Cheung and Rensvold (2002) first provided guidelines for acceptable invariance model fit, they did not examine various levels of non-invariance. Chen (2007) extended Cheung and Rensold's (2002) research and found significant influence of the model fit results under different levels of lack of invariance. However, recommendations for the goodness of fit statistics provided by these studies are based on continuous data using the maximum likelihood (ML) estimator. It is difficult to ascertain the validity of these prescriptions when using the weighted least squares mean and variance (WLSMV) estimator for ordered categorical data. Rutkowski and Svetina (2014) and Rutkowski and Svetina (2017) did investigate the performance of fit statistics with categorical indicators, but they did not specify which estimator they used and did not report the change of SRMR fit statistics.

To fill these gaps, this dissertation systematically investigated the sensitivity of changes in fit indices to the measurement invariance under various simulated conditions when ordered categorical data are used. Specifically, this study applied multiple-group CFA for ordered-categorical variables with a threshold structure. Two Monte Carlo studies were conducted to investigate three commonly used fit indices (CFI, RMSEA, SRMR) under conditions including different sample sizes, the number of indicators, source of non-invariance, levels of threshold symmetry, and proportion of non-invariant

items. Based on the simulation results, cutoff points were proposed for different levels of invariance. Rejection rates based on cutoff points of fit indices were discussed. Finally, the effects of a number of conditions were examined to determine the sensitivity of fit indices' changes to non-invariance. Violations of invariance based on cutoff points of goodness were discussed.

CHAPTER 3

METHOD

The current study, inspired by Chen (2007)'s work, was designed to fill the gaps in the literature remaining for evaluation of measurement invariance when ordered categorical data are analyzed. The general goal of this study was to examine the sensitivity of changes on three model fit indices (ΔCFI, ΔRMSEA, and ΔSRMR). To achieve this goal, two Monte Carlo simulation studies were conducted. Study 1 investigated the sampling variability of targeted fit indices under different levels of invariance including factor loadings, thresholds, residual variances, latent means, factor variances, and factor covariances. Based on the sampling variability of targeted fit indices, cutoff points for various non-invariance levels were preliminarily proposed. The goal of Study 2 was to investigate the influence of a number of conditions on the sensitivity of fit changes to two commonly used non-invariance levels encountered in empirical research: metric invariance where factor loadings are set to be equal across groups, and scalar invariance where thresholds are constrained to be equal across groups. Then, rejection rates based on the proposed cutoff points were examined in Study 2. All data were generated and analyzed using M*plus* software package (v. 8.6; Muthén & Muthén, 1998-2017), and *R* software package (R Core Team, 2020). Examples of simulation codes to generate and analyze study data were presented in Appendix A.

3.1 Study Design

3.1.1 Population Model

To achieve study goals, a two-group CFA model with ordered categorical indicators was designed for data generation in both Study 1 and Study 2. As shown in Figure 3.1, assuming simple structure, the number of factors in the population CFA model was fixed to two (Factor 1_k and Factor 2_k), and the number of groups (K) was restricted to two as well. The model included ten indicators, with an equal number of indicators (five) loaded on each factor. One group was treated as the reference group, and the other group served as the focal group.

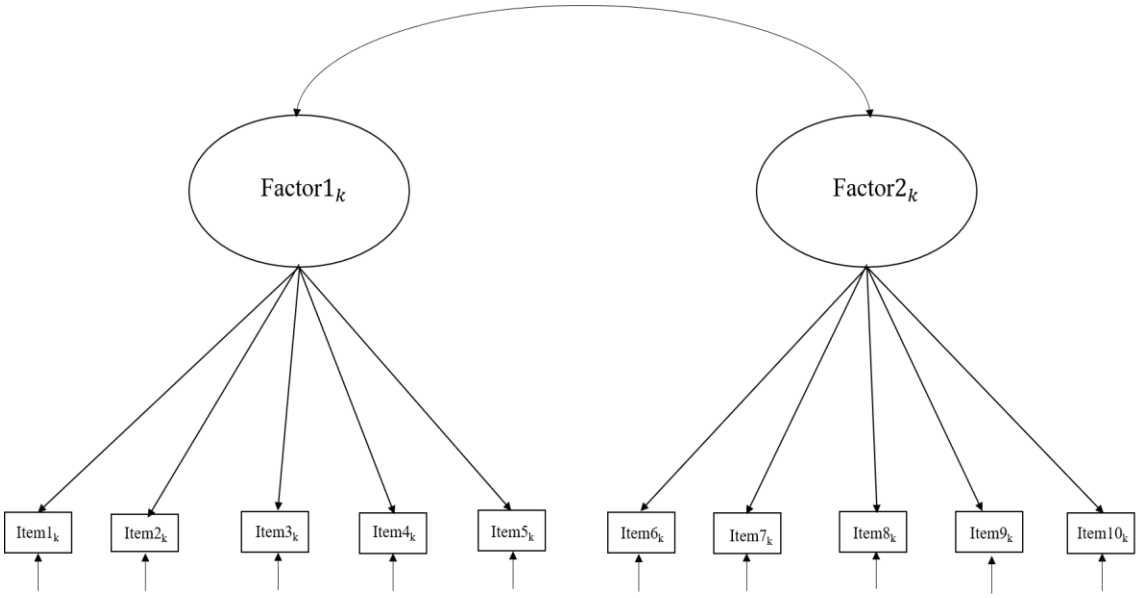

Figure 3.1 Population confirmatory factor analysis model

The number of factors and groups was restricted to two for simplicity, as well as for remaining practical when conducting a large simulation study (Hu & Bentler, 1998). Five indicators were selected because recommendations from simulation studies suggest that a minimum of three observed variables is needed for estimation and model identification purposes (Raykov & Marcoulides, 2012). In addition, Wolf et al. (2013)

suggested researchers may include more indicators per factor than the minimum required in order to compensate for small sample size and preserve statistical power. Therefore, to meet these criteria and to reflect typical context using CFA models (e.g., DiStefano & Hess, 2005; DiStefano et al., 2018), the number of items per factor was fixed to five in this study. One group was treated as the reference group where the factor mean and factor variance were set to be zero and one, and the factor mean and variance in the other group, which was the focal group, were set to be zero and one as well.

As shown in Table 3.1, all item loading values in both groups were held constant at a strong standardized loading size of 0.8 (Comrey & Lee, 2013; Wolf et al., 2013). Wolf et al. (2013) found that stronger factor loadings (e.g., 0.8) required smaller samples and had fewer problems with statistical power compared to the weaker factor loadings (e.g., 0.5). In addition, strong factor loadings without cross-loading may indicate good convergent validity (Cabrera-Nguyen, 2010). As a result, a loading size of 0.8 was selected in the population model, and a cross-loading condition was excluded from the population model. The population values for all residual variances were set to be 0.36. Finally, the factor correlation value was fixed to be 0.6 in the CFA model, which indicated a strong relationship between factors (r = 0.6).

3.1.2 Data & Data Generation

A small and medium sample size of 150 and 300 per group was generated in Study 1, resulting in a total sample size of 300 and 600. Data were generated with five categories using thresholds to denote data arising from a severe asymmetric observed distribution for all items (Rhemtulla et al., 2012). The selection of five categories in the

first study was based on a review of a set of published simulation studies examining SEM models with ordinal data (e.g., DiStefano et al., 2018; Flora & Curran, 2004).

The extreme asymmetric distribution was chosen because previous research has found that most methods performed worse when category thresholds were asymmetric (Rhemtulla et al., 2012). While the data were asymmetric at the observed level, the simulated data were assumed to come from an underlying normal distribution. The reason for using underlying normal distribution was that difference of observed data distributions was small when comparing an underlying normal distribution to an underlying nonnormal distribution (see supplemental document, Rhemtulla et al., 2012).

In summary, the population threshold values in the present study were set to be -1.34, -0.84, -0.44, and -0.05, resulting in 9%, 11%, 13%, 15%, and 52% of normally distributed data falling into each category (see details in Rhemtulla et al., 2012). Data properties of the population model were presented in Table 3.1.

Table 3.1 Population parameters for simulation Study 1

Parameter Controlled	Population values
Number of groups	2
Sample size per group	150 & 300
Number of factors	2
Number of items per factor	5
Number of ordered categories	5
Magnitude of factor loadings	0.8
Magnitude of item thresholds	-1.34, -0.84, -0.44, and -0.05
Factor means	0 and 0
Factor variances	1 and 1
Factor correlation	0.6
Estimator	WLSMV
Threshold symmetry	Extreme asymmetric
Underlying distribution	Normal

3.1.3 Estimation

The weighted least squares mean and variance adjusted (WLSMV) estimation method was used for data analysis. WLSMV is termed a "robust technique" meaning that it applies a correction to the original diagonally weight least square (DWLS) formula (DiStefano & Morgan, 2014). This estimation method was selected because the superior performance of WLSMV in estimating non-normal ordinal data has been well-established in the literature (Beauducel & Herzberg, 2006).

3.1.4 Study 1

Once the population model was set and the data were generated, a baseline CFA invariance model was fitted to the generated data. After estimating the baseline model, more restricted invariance models with a sequence of imposed equality constraints were fit into the generated data. The adequacy of a series of restrictions was compared based on the differences of model fit results. Specifically, the sequence started with testing configural invariance as the baseline. Once the model fit was adequate, metric invariance (equality constraints for factor loading across groups), scalar invariance (equality constraints for threshold across groups), factor variances, factor covariances, and factor means invariance were evaluated by calculating the differences of model fit indices on CFI, RMSEA, and SRMR.

Based on the differences of fit results among nested models, two outcomes were reported in Study 1: 1) sampling variability of changes in model fit across different levels of invariance; and 2) cutoff points for changes in model fit indices at different levels of invariance. A total of 1000 replications were conducted, resulting in a total of 2000 population datasets.

Considering model identification issue, residual variances were fixed to one when using WLSMV estimator, factor variances and factor means were fixed to the values defined in the population models (see details in Hoffman, n.d.; Millsap & Yun-Tein, 2004; Muthén &Asparouhov, 2002). In addition, to test for equal residual variances across groups, a backward invariance test procedure was proceeded in Study 1 (see detail steps in Hoffman, n.d), and theta parameterization was used for model specification.

At the test of configural invariance level, the fit of CFI, RMSEA, and SRMR were assessed using criteria of no smaller than 0.95 for CFI, and no larger than 0.05 for RMSEA and SRMR. When assessing factor loadings, thresholds, residual variances, factor variances, factor covariances, and latent means, changes of fit indices (ΔCFI, ΔRMSEA, and ΔSRMR) were obtained by calculating difference between the more restricted model and the baseline model.

3.1.5 Study 2

For Study 2, rejection rates for each change in model fit under a number of conditions were examined. An additional objective was to investigate the influence of the conditions on the sensitivity of changes on fit indices. Specifically, the same two-factor CFA population model was used when examining metric and scalar invariance, and five major conditions were manipulated: 1) sample size per group, 2) number of indicators, 3) source of non-invariance, 4) proportion of non-invariant items, and 5) threshold variability. These conditions were chosen based on prior simulation designs from methodological studies indicating their possibilities to affect model fit indices when measurement invariance is tested (Chen, 2007; Kim, 2012; Sass et al., 2014; Shi, 2016; Short, 2014). A summary of the studied conditions was presented in Table 3.2.

Table 3.2 Summary of conditions for simulation Study 2

Parameter Controlled	Condition Options
N (per group)	100, 600
Number of indicators	8, 16
Indicator categories	4
Source of non-invariance	Factor loadings only or thresholds only
Magnitude of invariant item loadings	0.8
Magnitude of non-invariant item loadings	0.5
Magnitude of invariant item thresholds	-1.25, 0, 1.25
Magnitude of non-invariant item thresholds	-1.23, -0.71, -0.28
Levels of threshold symmetry	Symmetry or extreme asymmetry
Proportion of non-invariant items	25% or 50%

Sample size. To achieve accurate estimates and ensure that CFA models can converge successfully, the sample size needs to be considered. Studies have found that most of the psychological research has followed an ad hoc rule of thumb requiring an N:p ratio of 10:1 in setting a lower bound for the sample sizes (Nunnally, 1967). Other researchers argue that a minimum sample size of 100 or 200 cases is preferable for structural equation modeling (SEM) (Boomsma, 1985), or 5 or 10 observations per estimated parameter is recommended (Bollen, 1989). Based on the empirical review study conducted by DiStefano and Hess (2005), the median sample size for empirical research was 377 across 101 reviewed articles, and only 19% of studies using CFA models considered with a sample size of less than 200 cases. A similar conclusion was also found in Jackson et al.'s (2009) study, indicating that over 90% of reviewed studies used adequate sample sizes, with only 7.7% of studies used very small samples less than 100. Overall, the rule of N:p ratio of 10:1 is still widely used and satisfied in a majority of applied research studies.

Although most studies used a large sample size, many applied researchers and practitioners argued that sometimes they were unable to obtain adequate sample sizes due

to budget constraint or limited target population (Nevitt & Hancock, 2004; Westland, 2010). Considering a small sample size condition in the simulation study may benefit applied researchers, as well as reflect a common dilemma in empirical research, two conditions were included in this study to emphasize possible limitations in applied research. Specifically, the sample sizes were set to be 100 (a small sample size), and 600 (a large sample size) for each group with a N:p ratio of 10:1, and 60:1. A medium sample size (N= 300 per group) was not included in Study 2 because it had been considered in Study 1. The total sample sizes were 200, and 1200 with sample sizes equally distributed in both groups.

Number of indicators. The number of indicators used in previous simulation studies with CFA models varied widely. For example, DiStefano and colleagues (2018) simulated a commonly applied three-factor CFA model including a total number of 15 items with five items per factor. Flora and Curran (2004) tested four models with five to 10 indicators per factor. Shi et al. (2019) examined the influence of model size on SEM fit indices by simulating a two-factor CFA model with a total number of observed variables ranging from 10 to 120. In addition, DiStefano et al. (2019) employed a three-factor CFA model with five to 60 items per factor. Based on the review study, DiStefano and Hess (2005) noted that the medium model size used in applied research studies was found to be a four-factor model with 16 indicators, approximately 4-7 indicators per factor. Jackson et al. (2009) reviewed 194 published studies and found that the median number of observed variables in the models was 17, with 25% of the studies using models less than 12 variables and with 25% of the studies using models more than 24 items. Therefore, based on existing literature, the number of indicators for each factor in

this study were set to be 5 and 8 yielding the total number of indicators of 10 (2 factors/10 indicators) and 16 (2 factors/16 indicators). These two conditions covered a small and medium model size, which approximate conditions used in most previous CFA studies. A large model size (i.e., 8 factors/40 indicators) was not considered as it appears to be less common in applied research.

Indicator categories. This condition specified how many categories were included for each indicator. As documented in many empirical studies, variables characterized by an ordinal scale of measurement (i.e., Likert-type items) are common within social and behavioral sciences (Flora & Curran, 2004). Based on a set of previous articles, categories less than five are typically used in empirical studies to investigate issues related to ordered categorical data (i.e., DiStefano et al., 2018; Flora & Curran, 2004; Sass et al., 2014). Commonly, five-category is used as a cutoff for defining ordered data as items with more than five categories are often treated as continuous variables, thus, can be estimated using maximum likelihood (ML) estimation (Kim, 2012). In Study 2, four categories were selected for comparison with Study 1 results with five categories. Further, four-category scale is the most ideal in many real-world situations when researchers want to exclude participants' neutral answer (Chyung et al., 2017).

Usually, ordered categorical variables are generated by categorizing continuous variables, where the underlying distribution is unknown, but is assumed to be normally distributed (DiStefano et al., 2018; Rhemtulla, et al., 2012). Lubke and Muthén (2004) noted that while ordered categorical data with non-normal distributions have been extensively investigated in single-group models, studies on the performance of model fit indices using MLR or WLSMV estimators with ordered categorical data for multiple-

group models are rare. To be consistent with previous simulation designs and to add novel findings in the literature, this study chose an underlying normal distribution as a simulated condition.

Source of non-invariance. The source of non-invariance can vary at different places for parameters under consideration with invariance testing. For example, Sass et al. (2014) designed three non-invariant locations in their Monto Carlo simulation study: factor loadings only, thresholds only, and both factor loadings and thresholds, to evaluate the performance of metric, scalar, and cumulative non-invariance. Their findings indicated that the source of non-invariance can substantially impact the power of Δ chi-square when using different estimators, including ML, MLR, and WLSMV (Sass et al., 2014). Both Kim (2011) and Shi (2016) varied the locations of non-invariance at either factor loadings or intercepts/thresholds to examine the influence of sources of non-invariances on the targeted models. According to Kim (2011), different measurement invariance testing techniques (e.g., multiple-group CFA) may be employed depending on the source of non-invariance.

As many investigators have found the source of non-invariance may potentially impact research results, adding this factor into a simulation study is necessary. In this study, the location of non-invariance was manipulated either on factor loadings or thresholds without simulating both at the same time. These two conditions were used to evaluate the performance of metric and scalar measurement non-invariance.

When testing lack of loading invariance (metric measurement non-invariance), a weaker loading size was used for the non-invariant items in Study 2. Specifically, the standardized factor loading size for the invariant items were fixed to 0.8, and the

standardized factor loading size for the non-invariant items was designed to be 0.5 (a moderate loading size) (e.g., DiStefano et al., 2018; Sass et al. 2014; Shi, 2016).

When testing lack of thresholds invariance (scalar measurement non-invariance), the invariant item threshold values were set to be 1.25, 0, 1.25, which indicated a symmetric threshold condition, and non-invariant item threshold values were fixed to be 1.23, -0.71, and -0.28, representing an extreme asymmetric threshold condition.

Levels of threshold symmetry. With respect to thresholds, two conditions were included in the simulation study: 1) symmetry condition; and 2) extreme asymmetry condition. Threshold symmetry values were adapted based on suggestions from previous literature (Rhemtulla, et al., 2012). In the symmetry condition, the underlying normal distribution is evenly discretized through a set of threshold values that are represented by Z-scores. Specifically, for four categories, threshold values were set to be -1.25, 0, 1.25, resulting in 11%, 39%, 39%, and 11% of normally distributed data falling into each category. In the extreme asymmetry condition, category threshold values were created so that the peak of the distribution fell to the right of the center. Specifically, the category threshold values were -1.23, -0.71, and -0.28 for four-category, which resulted in 11%, 13%, 15%, and 61% of normally distributed fata falling into four categories. A summary of threshold values used in the study 2 was shown in Table 3.3.

Table 3.3 Threshold conditions

Threshold condition	No. cat	Threshold values	Proportion of cases in each category			
Symmetry	4	-1.25, 0, 1.25	11	39	39	11
Ext. Asym	4	-1.23, -0.71, -0.28	11	13	15	61

Proportion of non-invariant items. Finally, the proportion of non-invariant items was considered, as prior research has found that this factor affected the changes in model fit indices to lack of measurement invariance when ML estimation with continuous data was used (Cheng, 2007). Two conditions were considered in this study: 25% of the non-invariant items, and 50% of the non-invariant items. The 25% of the non-invariant items indicated the low contamination situation (Shi, 2016), where only 25% of item loadings were different across the two groups. Similarly, 50% of the non-invariant items represented a high contamination condition (Shi, 2016), where 50% of item loadings were different across the two groups. These proportions were selected based on previous simulation research studies for testing measurement invariance (Shi, 2016; French & Finch, 2008).

In total, this simulation study consisted of 32 fully crossed conditions: 2 levels of sample sizes (100, and 600 per group) * 2 level of the number of indicators (4, 8 per factor) * 1 level of category condition (four-category) * 2 levels of source of non-invariance (factor loadings or thresholds) * 1 level of magnitude of item loadings (item loadings of 0.8 for invariant items and item loadings of 0.5 for non-invariant items) * 2 levels of threshold variability (symmetry, and extreme asymmetry) * 2 levels of the proportion of non-invariance on items (25%, and 50%). All data were generated and analyzed using a multiple-group CFA model. For each designed simulation condition, one thousand replications were run. Replications that exhibited non-convergence or improper solutions were removed and only results converging to a proper solution was included in the analyses.

3.2 Procedures and Analyses

All the data generation, estimation and analyses were conducted using M*plus* version 8.6 (Muthén & Muthén, 2017) and *R* (R Core Team, 2020). In Study 1, the sampling variability of model fit indices under different levels of invariance was summarized by tables across conditions and invariance levels. Descriptive results were given including the means, standard deviations, 1^{st} and 5^{th} percentiles of changes in CFI, as well as the means, standard deviations, 95^{th} and 99^{th} percentiles of changes in RMSEA and SRMR. Based on the results of Study 1, cutoff values were given for testing measurement invariance at levels of factor loadings, thresholds, residual variances, factor variances, factor covariances, and factor means. The proposed cutoff points were based on the average value of the means, $1^{st}/95^{th}$ or $5^{th}/99^{th}$ percentiles in fit differences.

In Study 2, five major factors were considered: sample size, number of indicators, source of non-invariance, levels of threshold symmetry, and proportion of non-invariant items. Results of rejection rates based on cutoff points of changes in fit indices were examined across studied conditions.

In summary, this dissertation included two Monto Carlo simulation studies to investigate the performance of changes in three model fit indices with ordered categorical data in the context of measurement invariance testing with multiple-group CFA under various conditions commonly founded in both applied and methodological studies. The findings may contribute to previous research and provide both applied and simulation researchers with a baseline reference as to how changes in model fit perform under the simulated modeling conditions.

Chapter 4 presented the results of these two studies. In chapter 4, model convergence and performance of all ad-hoc model fit indices under all simulated conditions were examined. Cutoff guidelines for the changes in model fit indices to the multiple-group CFA invariance testing were discussed. Rejection rates throughout all conditions were investigated. Finally, the impact of model characteristics was described throughout this chapter.

CHAPTER 4

RESULTS

This study investigated the performance of model fit indices when testing measurement invariance in the context of multiple-group CFA. The study used ordinal data with the analyses and the research design included simulated conditions commonly encountered in practice. Two Monte Carlo studies were designed to answer four research questions.

Study 1 was designed to examine sampling variability of fit indices under population conditions, assuming invariance across test levels. Two research questions were addressed with this study. Research Question 1 examined the sampling variability of three fit indices under various invariance levels including factor loadings, thresholds, residual variances, latent means, factor variances, and factor covariances, with a goal of providing applied researchers assistance when evaluating measurement invariance with ordinal data. Research Question 2 examined whether the proposed criteria of changes in model fit indices were consistent with each level of invariance including successively restricting factor loadings, thresholds, residual variances, latent means, factor variances, and factor covariances.

Study 2 aimed to examine the performance of fit indices across different conditions when testing two common non-invariance levels: metric non-invariance and scalar non-invariance. This study addressed Research Questions 3 and 4. Research Question 3 aimed to investigate the influence of various simulated conditions on the

performance of changes in fit indices under two commonly used non-invariance levels: metric non-invariance and scalar non-invariance. Then, rejection rates based on cutoff points of fit indices proposed in Study 1 were examined in Study 2. Last, Research Question 4 compared the proposed standards of ordinal data with fit criteria commonly used with invariance evaluation when continuous data are analyzed.

4.1 Performance of Model Fit Indices for Study 1.

The first study followed the earlier work of Cheung and Rensvold (2002) as well as Chen (2007). Overall, data were simulated across various conditions, assuming the null hypothesis of invariance. Convergence rates in Study 1 were examined to assess the percentage of successfully converged replications for each simulated condition. All models successfully converged one thousand times.

Table 4.1 displayed the means, standard deviations, and the 1^{st} and 5^{th} percentiles of CFI, and 95^{th} and 99^{th} percentiles of RMSEA and SRMR for sample sizes of 300 and 600, respectively. Table 4.2 presented the means, standard deviations, and 1^{st} and 5^{th} percentiles of ΔCFI, and 95^{th} and 99^{th} percentiles of ΔRMSEA, and ΔSRMR for both sample size conditions. The percentiles shown in the tables indicated various critical values to use for rejecting the null hypothesis of invariance. The results were discussed by two types of tests: measurement invariance tests including invariance of model form (configural invariance), factor loadings (metric invariance), thresholds (scalar invariance), and residual variances (strict invariance), and structural invariance tests including tests of factor variances, factor covariances, and factor means invariance.

First, as the sample increased, the performance of fit results improved and the sampling variation in fit indices decreased. For example, when testing factor loading

invariance, as the sample increased from 300 to 600, the means of RMSEA decreased from 0.016 to 0.011, and the associated standard deviations decreased from 0.017 to 0.012. The 95th percentiles for RMSEA decreased from 0.046 to 0.033. A similar pattern was detected for the CFI and SRMR in terms of the means, percentiles, and standard deviations.

Second, when testing configural invariance, as expected, all three fit indices supported the hypothesis of equal form across groups at both sample sizes 300 and 600. For example, mean values of CFI were 0.998, and 1.000 for sample sizes 300 and 600, respectively, and the 5th percentiles of CFI were 0.991, and 0.995. The mean values of RMSEA were 0.015, and 0.010, and the 95th percentiles of RMSEA were 0.046, and 0.033 for both conditions. Mean values of SRMR were 0.035, and 0.024, and the 95th percentiles of SRMR were 0.042, and 0.029 with sample sizes of 300 and 600. Overall, all simulated results were consistent with the population models for testing equal model form.

Third, when testing measurement invariance in loadings, thresholds, and residual variances with both sample sizes 300 and 600, all three fit indices (CFI, RMSEA, and SRMR) were more sensitive to random variation in factor thresholds and residual variances than in factor loadings, while changes in SRMR were relatively smaller across thresholds and residual variances. For example, given a model with a sample size of 600, the 95th percentiles of ΔRMSEA were 0.00, 0.043, and 0.048 for invariance tests of loadings, thresholds, and residual variances, respectively. For ΔCFI, the 5th percentiles were -0.001, -0.037, and -0.037; For ΔSRMR, the 95th percentiles were 0.004, 0.10, 0.12 for invariance tests of loadings, thresholds, and residual variances, respectively.

Figures 4.1 and 4.2 reported the fit differences of $5^{th}/95^{th}$ percentiles based on loadings, thresholds, and residual variances invariance tests for sample size 300 and 600. CFI is an incremental fit index that compares a hypothesized model fit with a model with the worst fit. Therefore, larger values demonstrate "better" model fit. On the contrary, RMSEA and SRMR indicate the "badness" of model fit that assesses how far a hypothesized model is from a perfectly fitting model. As a result, smaller values are used to show how much "better" the model fits as compared to the true value. Note that the pattern for CFI is, by definition, opposite to the patterns of RMSEA and SRMR.

Figures 4.1 and 4.2 showed that among the three fit indices, SRMR was slightly more sensitive to random variation in factor loadings than CFI and RMSEA for both sample sizes 300 and 600, as the changes in SRMR was found to be larger than the changes in CFI and RMSEA. However, when sample size increased to 600, the sensitivity of SRMR to random variation in factor loadings was not obvious, as change in SRMR decreased slightly.

Last, instead of producing all positive ΔRMSEA values, the study also identified a negative 99^{th} percentile value (-0.001) in the 300-sample size condition when testing loading invariance. Although this result was not expected, as the more constrained model should perform less well than the less constrained model, previous invariance research studies have yielded similar findings (e.g., Rutkowski & Svetina, 2014; 2017).

Fourth, when testing structural invariance in factor variances, covariances, and means for sample size 300, all fit indices were more sensitive to random variation in factor variances than covariances and latent mean values. For example, for CFA, the 5^{th} percentiles were -0.033, 0.000, and 0.020 for invariance tests of factor variances,

covariances, and the latent means, respectively (see Figures 4.3 and 4.4 for 5th/95th percentiles). For RMSEA, the 95th percentiles were 0.027, 0.000, and -0.020; For SRMR, the 95th percentiles were 0.028, 0.018, and -0.020. A similar pattern was also detected in percentiles when the sample size increased to 600. Additionally, compared to ΔCFI and ΔRMSEA with sample size 300, ΔSRMR was the least sensitive to random variation in latent means.

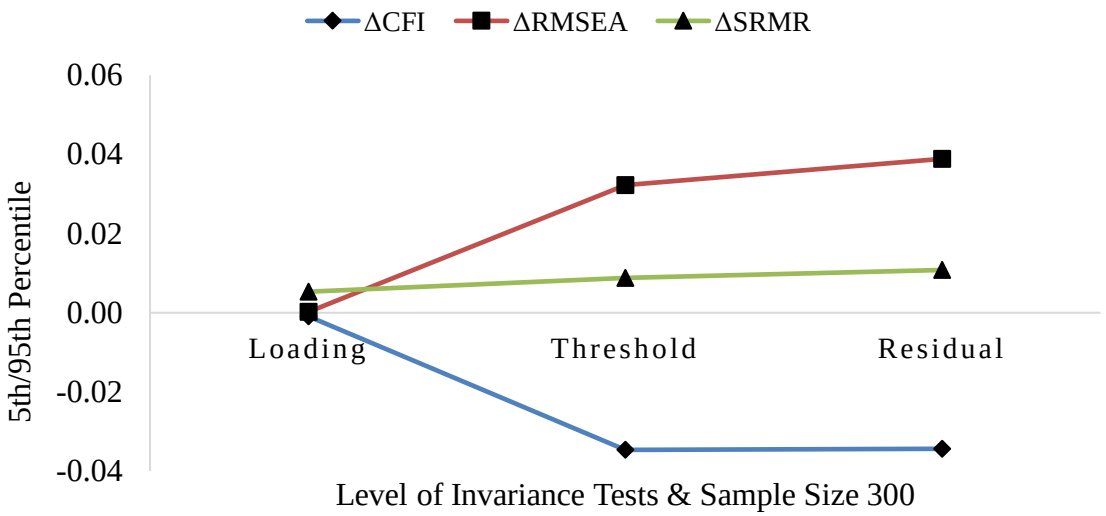

Figure 4.1 The 5th/95th percentile of differences in fit indices based on different levels of measurement invariance tests for sample size 300. Note: the pattern of CFI was opposite to the patterns of RMSEA and SRMR.

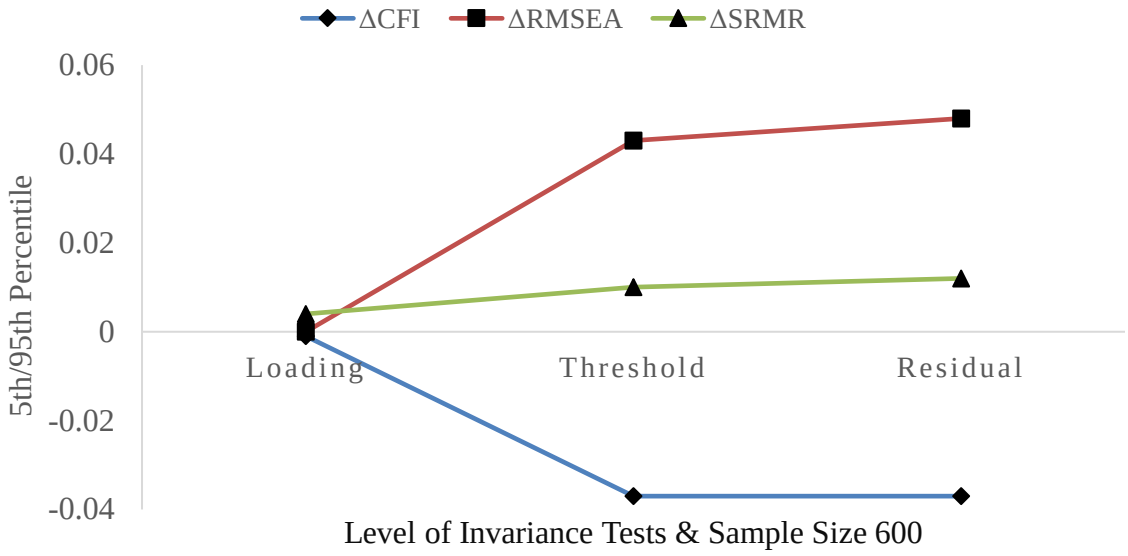

Figure 4.2 The 5th/95th percentile of differences in fit indices based on different levels of measurement invariance tests for sample size 600. Note: the pattern of CFI was opposite to the patterns of RMSEA and SRMR.

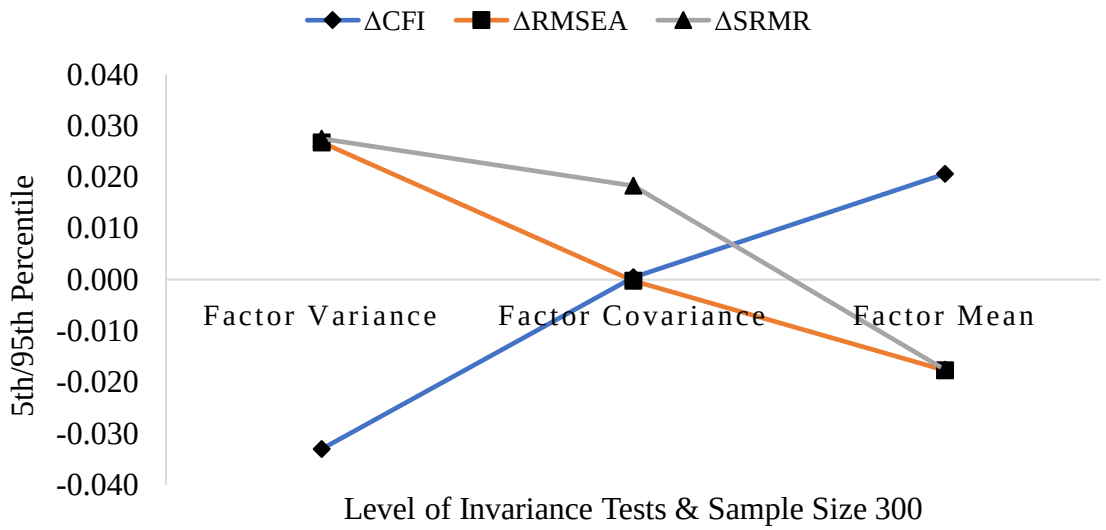

Figure 4.3 The 5th/95th percentile of differences in fit indices based on different levels of structural invariance tests for sample size 300.

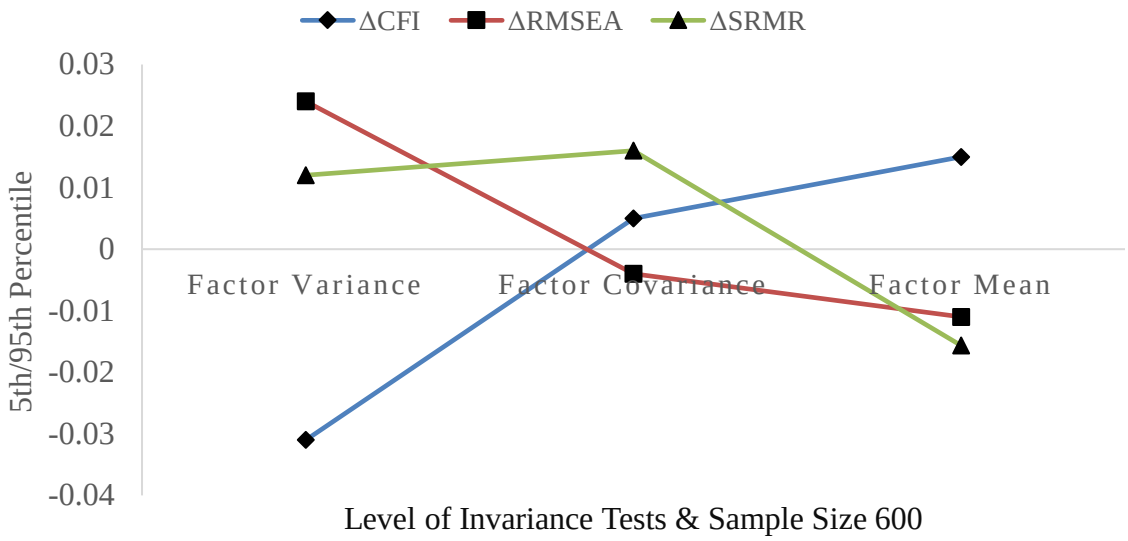

Figure 4.4 The $5^{th}/95^{th}$ percentile of differences in fit indices based on different levels of structural invariance tests for sample size 600.

Table 4.1 Goodness of fit indices under different levels of invariance

	CFI				RMSEA				SRMR			
N	M	SD	5%	1%	M	SD	95%	99%	M	SD	95%	99%
Configural Invariance												
300	0.998	0.003	0.991	0.988	0.015	0.017	0.046	0.055	0.035	0.004	0.042	0.045
600	1.000	0.002	0.995	0.993	0.010	0.012	0.033	0.039	0.024	0.003	0.029	0.031
Metric Invariance												
300	0.998	0.003	0.990	0.987	0.016	0.017	0.046	0.053	0.039	0.005	0.047	0.051
600	0.999	0.002	0.995	0.992	0.011	0.012	0.033	0.040	0.028	0.003	0.033	0.035
Scalar Invariance												
300	0.971	0.009	0.956	0.947	0.064	0.009	0.078	0.085	0.049	0.004	0.056	0.059
600	0.968	0.006	0.957	0.952	0.067	0.006	0.076	0.080	0.039	0.003	0.043	0.045
Strict Invariance (Model A)												
300	0.998	0.004	0.990	0.985	0.013	0.015	0.039	0.048	0.038	0.004	0.045	0.048
600	0.999	0.002	0.995	0.992	0.009	0.010	0.028	0.035	0.027	0.003	0.032	0.033
Strict Invariance (Model B)												
300	0.971	0.009	0.956	0.947	0.064	0.009	0.078	0.850	0.049	0.004	0.056	0.059
600	0.968	0.006	0.957	0.952	0.067	0.006	0.076	0.080	0.039	0.003	0.043	0.045
Factor Variance Invariance												
300	0.953	0.017	0.923	0.907	0.081	0.014	0.105	0.113	0.068	0.009	0.084	0.091
600	0.948	0.013	0.927	0.914	0.084	0.010	0.100	0.107	0.060	0.007	0.071	0.078
Factor Covariance Invariance												
300	0.958	0.018	0.923	0.905	0.076	0.017	0.105	0.120	0.078	0.013	0.101	0.113
600	0.956	0.013	0.932	0.921	0.077	0.011	0.095	0.103	0.069	0.010	0.087	0.094
Factor Mean Invariance												
300	0.967	0.013	0.944	0.930	0.068	0.012	0.087	0.096	0.069	0.009	0.084	0.091
600	0.964	0.009	0.948	0.939	0.069	0.008	0.082	0.089	0.061	0.007	0.072	0.079

Note. Comparison of equal residual variances was conducted backward, meaning that strict invariance (Model A) with all residual variances freely estimated in the second group was fitted first, and then compared with strict invariance (Model B) with all residual variances fixed to the population value (0.36) in the second group.

Table 4.2 Sampling variability of goodness of fit indices under different levels of invariance

	ΔCFI				ΔRMSEA				ΔSRMR			
N	M	SD	5%	1%	M	SD	95%	99%	M	SD	95%	99%
Loading Invariance (Baseline: Configural)												
300	0.000	0.002	-0.005	-0.006	0.001	0.012	0.000	-0.001	0.005	0.002	0.005	0.006
600	0.000	0.001	-0.001	-0.001	0.001	0.008	0.000	0.000	0.003	0.002	0.004	0.004
Threshold Invariance (Baseline: Metric)												
300	-0.026	0.008	-0.035	-0.040	0.048	0.014	0.032	0.032	0.010	0.002	0.009	0.008
600	-0.031	0.006	-0.037	-0.041	0.056	0.011	0.043	0.040	0.011	0.002	0.010	0.010
Residual Invariance (Baseline: StrictA)												
300	-0.026	0.008	-0.034	-0.039	0.051	0.012	0.039	0.037	0.011	0.002	0.011	0.010
600	-0.031	0.006	-0.037	-0.041	0.058	0.010	0.048	0.045	0.012	0.002	0.012	0.012
Factor Variance Invariance (Baseline: StrictB)												
300	-0.019	0.014	-0.033	-0.040	0.017	0.013	0.027	0.028	0.019	0.008	0.028	0.032
600	-0.020	0.011	-0.031	-0.037	0.017	0.009	0.024	0.027	0.012	0.002	0.012	0.012
Factor Covariance Invariance (Baseline: Factor Variance)												
300	0.005	0.008	0.000	-0.002	-0.006	0.007	0.000	0.007	0.009	0.008	0.018	0.023
600	0.008	0.006	0.005	0.006	-0.007	0.005	-0.004	-0.005	0.008	0.006	0.016	0.015
Factor Mean Invariance (Baseline: Factor Covariance)												
300	0.010	0.010	0.020	0.030	-0.010	0.010	-0.020	-0.020	-0.010	0.010	-0.020	-0.020
600	0.010	0.010	0.020	0.020	-0.010	0.010	-0.010	-0.010	-0.010	0.010	-0.020	-0.010

Note. Comparison of equal residual variances was conducted backward, meaning that strict invariance (Model A) with all residual variances freely estimated in the second group was fitted first, and then compared with strict invariance (Model B) with all residual variances fixed to the population value (0.36) in the second group.

Based on the results of Study 1, cutoff criteria were proposed for testing measurement invariance at metric invariance (factor loadings), scalar invariance (thresholds), and strict invariance (residual variances) levels. These levels were chosen to align with the commonly evaluated tests for measurement invariance in practice. Following Chen's (2007) guideline, the proposed cutoff points are roughly based on the mean values in changes at the $1^{st}/95^{th}$ or $5^{th}/99^{th}$ percentiles of fit under the null hypothesis that a given level of invariance holds across two sample size conditions. Tables 4.1 and 4.2 showed that sample size impacts sampling variation in changes of fit indices, where sampling variation increases as the sample size decreases. Meanwhile, according to Chen's study (2007), it is easier to commit Type I errors (i.e., the probability of rejecting the null hypothesis when it is true) when the sample size is small and to commit Type II errors (the probability of accepting the null hypothesis when it is false) when the sample size is large. Therefore, an adequate cutoff criterion should minimize both Type I and II errors at the same time (Hu & Bentler, 1999). Overall, cutoff criteria in this study are proposed considering the influence of sample size.

When comparing the configural model with the metric model, and the metric model with the scalar model, the results indicated an increasing sensitivity to random variation in factor thresholds and residuals rather than factor loadings. Thus, different cutoff points were recommended for different levels of invariance tests: when testing loading invariance, a change of $\leq |\pm 0.003|$ is proposed because the average value of ΔCFIs across means, 1^{st} and 5^{th} percentiles of ΔCFIs was around $|\pm 0.003|$. However, when testing threshold and residual variance invariance levels, a change of $\leq |\pm 0.03|$ is recommended considering that the average value of ΔCFIs across means, 1^{st} and 5^{th}

percentiles of ΔCFIs was around |±0.03|. It is noted that absolute cutoff values are recommended because positive ΔCFIs were found in many conditions. For example, when testing item thresholds (scalar) invariance, 32.2% ΔCFIs across 1000 replications were positive in condition with a model sample size of 100 per group, 8 indicators, extreme asymmetric, and 50% non-invariant item thresholds. Another reason for using absolute cutoff values is that this study only focused on investigating magnitude of fit changes across different levels of invariance rather than fit improvement.

Similar to CFA, RMSEA was also more sensitive to random variation in factor thresholds and residuals than factor loadings. Therefore, two cutoff points were proposed: when testing factor loading invariance, a change of ≤ |±0.001| can be used, especially when the sample size is smaller than 300; when testing threshold and residual variance invariance, a change of ≤ |±0.02| is recommended.

Last, for SRMR, the same value is suggested for all three levels of invariance given that SRMR was almost equally sensitive to all three levels of invariance especially when the sample size is small (e.g., 300): a change of ≤ 0.007 is proposed. These proposed cutoff values are applied to the next study to examine the rejection rates under various degrees of invariance.

4.2 Performance of Model Fit Indices for Study 2

Study 2 was conducted to investigate the effect of various simulated conditions on the performance of changes in fit indices under two commonly used invariance levels: metric invariance and scalar invariance. The second goal was to examine the rejection rates on cutoff points of fit indices proposed in Study 1. Last, the proposed standards for

invariance evaluation with ordinal data were compared with standards commonly applied with continuous data.

The data generation procedure was similar to the procedure used in Study 1, but five factors that might impact the changes of fit indices to invariance testing were considered: sample size in each group, number of indicators, source of non-invariance, levels of threshold symmetry, and proportion of non-invariance (see details in Chapter 3). A total of 32 fully crossed conditions were simulated, and one thousand replications were generated for each simulation condition. Any non-convergence or improper solutions were removed from the study, and the number of convergences was increased until reached to 1000 successful iterations.

To determine the effect of the simulated conditions on the model fit indices, descriptive information including means, standard deviations, $5^{th}/95^{th,}$ and $1^{st}/ 99^{th}$ percentiles of the three model fit indices were examined across all cells of the study. Then, average values were compared with the cutoff points proposed in Study 1. Specifically, when testing loading invariance, ΔCFI is equal or less than |±0.003|; ΔRMSEA is equal or less than |±0.001|; and ΔSRMR is equal or less than 0.007; when testing threshold and residual variance invariance levels, ΔCFI is equal or less than |±0.03|; ΔRMSEA is equal or less than |±0.02|, and ΔSRMR is equal or less than 0.007.

Overall convergence rates were high across all the simulated conditions (see Table 4.3). Convergence problems only occurred when invariance tests were examined at the lowest sample size data and with 50% non-invariant item thresholds. The lowest convergence rate was 94.6% across three levels of invariance tests. Two main reasons of non-convergence were noted: a non- positive definite latent variable covariance matrix

and a computation issue related to standard errors of the model parameter estimates. In summary, the high convergence rates indicated an adequate estimation of the model parameters.

4.2.1 CFI

Descriptive results of CFI differences were shown in Tables 4.4 and 4.5, Figures 4.5, 4.6, 4.7 and 4.8. Over all study conditions, there was very little variability observed within ΔCFI values across the metric and scalar invariance conditions. Thus, the ΔCFI was sensitive to the lack of invariance for both metric and scalar invariance tests.

Sample size. There was a slight increase in ΔCFIs when the sample size increased from 200 to 1200. Also, the sample size impacted standard deviation of changes in CFI significantly. For example, when testing factor loading invariance, given the 8-indicator model with symmetric thresholds and 25% non-invariant item loadings, the mean of ΔCFI values varied from -0.007 to -0.010 for sample sizes 200 and 1200. However, the standard deviations decreased from 0.010 to 0.003 across the span of sample sizes tested (200 to 1200). A similar pattern was also observed across study conditions when testing factor threshold invariance. Interestingly, however, when the non-invariant item loadings increased to 50% of total items or when the non-invariant thresholds were simulated in the models, changes slightly decreased for factor loading invariance. For example, given the 8-indicator model with symmetric thresholds and 50% non-invariant items, mean changes in CFI were -0.001 vs. 0.000 for both item loadings and thresholds between sample size 200 and sample size 1200. This pattern was not observed for item threshold invariance testing.

Number of indicators. The changes in CFI were larger for the 8-indicator models than in 16-indicator models when testing item threshold non-invariance; however, the pattern was opposite when factor loading non-invariance was examined. For example, given a model with symmetric thresholds, 25% non-invariant item loadings, and sample size 200, the mean values of ΔCFI changes were -0.008 with 8-indicator models (vs. -0.005 with 16-indicator models) for factor threshold invariance testing, whereas the means of ΔCFI were -0.007 vs. -0.011 for factor loading invariance testing. The standard deviations were higher in the 8-indicator conditions than the 16-indicator conditions.

Source of non-invariance. Two locations of non-invariant items were examined in this study: 1) non-invariant item loadings only, and 2) non-invariant item thresholds only. The changes in CFI were larger when testing factor threshold non-invariance than when testing factor loading non-invariance (see Tables 4.4 and 4.5). The results indicated that ΔCFI is more sensitive to the tests of factor thresholds than factor loadings across the study conditions. For example, given an 8-indicator model with symmetric thresholds, and sample size of 200, the means of CFI changes were -0.007 and -0.001 with 25% of non-invariant item loadings and 25 % of non-invariant item thresholds respectively when testing factor loading non-invariance (vs.-0.008 and -0.033 when testing factor threshold non-invariance), and the means of CFI changes were -0.001 and -0.001 with 50% of non-invariant item loadings and 50% of non-invariant item thresholds respectively when testing factor loading non-invariance (vs. -0.017 and -0.004 when testing factor threshold non-invariance).

Levels of threshold symmetry. The levels of threshold symmetry did not have an appreciable impact on changes in CFI when factor loading non-invariance was assessed.

However, the changes in CFI were larger in the symmetric threshold model than the extreme asymmetric threshold model when factor threshold non-invariance was tested, except for models with 50% non-invariant thresholds. These findings indicated that ΔCFI is more sensitive to detecting symmetric threshold non-invariance than extreme asymmetric threshold non-invariance. For example, given a 16-indicator model with sample size of 200 and 25% non-invariant item loadings, the means of CFI change were -0.011 for both symmetric and extreme asymmetric threshold conditions when testing factor loading non-invariance. However, the means of CFI changes were -0.005 for symmetric threshold condition vs. -0.002 for extreme asymmetric threshold condition when testing factor threshold invariance.

Proportion of non-invariant items. Concerning the proportion of non-invariant item loadings, changes in CFI were bigger with 25% non-invariant item loadings were included than when 50% non-invariant item loadings were present. The pattern was opposite for lack of threshold invariance. For example, given a 16-indicator model with sample size of 1200 and extreme asymmetric thresholds, the changes in CFI were -0.013 when there were 25% non-invariant item loadings (vs. 0.000 when the non-invariant item loadings were 50%). For lack of loading invariance, however, the changes in CFI were -0.002 vs. -0.004 for lack of threshold invariance with the same simulated condition.

Considering the proportion of non-invariant item thresholds, changes in CFI were small and consistent across the 32 study conditions when testing item loading invariance. This finding was expected, as no non-invariant item thresholds were simulated in the population models. The pattern of changes in CFI was inconsistent across different proportion conditions, but changes in CFI did increase when item threshold values were

not invariant. For example, given an 8-indicator model with symmetric item thresholds and sample size of 200, the changes in CFI were -0.033 when the non-invariant item thresholds were 25% vs. -0.004 when the non-invariant item thresholds were 50%. However, given the 8-indicator model with extreme asymmetric item thresholds and sample size of 200, the changes in CFI were -0.021 with 25% non-invariant item thresholds vs. -0.032 with the 50% non-invariant item thresholds.

Rejection Rates. Rejection rates for changes in CFI across sample sizes and levels of invariance are shown in Tables 4.10 and 4.11. Based on the review of literature with both continuous data and ordered data, as well as based on the results of Study 1, several cutoff values were examined: 1) |± 0.002| (Mead et al, 2008); 2) |± 0.003| from Study 1; 3) |± 0.03| from Study 1; and 4) -0.005 and -0.01 (Chen, 2007).

First, considering the impact of sample size, the results indicated that rejection rates based on ΔCFI appeared to vary across study conditions with different sample sizes. Specifically, when testing both factor loading and threshold non-invariance, rejection rates of ΔCFI tended to increase across 25% non-invariant item loading conditions as sample size increased. However, rejection rates of ΔCFI decreased substantially for sample size of 1200 when factor loading non-invariance was examined, especially when the models included 25% non-invariant item thresholds, as well as 50% non-invariant item loadings or thresholds. For example, using |± 0.002| as the cutoff value, given a 16-indicator model with symmetric items and 50% non-invariant item loadings, the rejection rates were 41.4% when the sample size was 200 vs. 0.7% when the sample size was 1200.

Second, for the number of indicators, it seems that rejection rates did not change significantly between 8-indicator models and 16-indicator models. For example, given a model with symmetric item thresholds, sample size of 200, and 25% non-invariant item thresholds for lack of threshold invariance, the rejection rates in CFI were 99.9% for the 8-indicator model vs. 99.4% for the 16-indicator model using a cutoff value of 0.002.

Third, in terms of source of non-invariance, as expected, the rejection rates were higher when the source of non-invariance was from item loadings than thresholds for lack of loading invariance; however, the rejections rates were inconsistent when the source of non-invariance was from item thresholds for lack of threshold invariance. For example, given an 8-indicator model with symmetric threshold, and sample size of 200 for lack of loading invariance, the rejection rates in CFI were 77.5% with 25% non-invariant loadings vs. 51.0% with 25% non-invariant item thresholds when testing factor loading non-invariance using 0.002 as a cutoff value. The rejection rates in CFI, which performed inconsistently, were 86.4% and 95.7% with 25% and 50% of non-invariant item loadings vs. 99.9% and 58.9% with 25 % and 50% of non-invariant item thresholds when testing threshold non-invariance using 0.002 as a cutoff value.

Fourth, the levels of threshold symmetry were not an impactful factor on the rejection rates for ΔCFI. For example, given a 16-indicator model with sample size of 200 and 25% non-invariant item loadings for lack of loading invariance, the rejection rates in CFI change were 95.2% vs. 89.5% for symmetric and extreme asymmetric threshold conditions when testing factor loading non-invariance using 0.002 as a cutoff value.

Last, considering the proportion of non-invariant items with both item loadings and item thresholds, 25% of non-invariant items had higher rejection rates than 50% non-invariant items in most study conditions. For example, given a 16-indicator model with sample size of 1200 and extreme asymmetric threshold for lack of item loading invariance, the rejection rates in CFI were 100.0% when there were 25% non-invariant item loadings vs. 2.7% when the non-invariant item loadings were 50% using 0.002 as a cutoff value.

Overall, on average, the CFI was more effective at identifying threshold non-invariance (scalar non-invariance) across 32 study conditions than loading non-invariance (metric non-invariance). The average CFI difference across the conditions and across two levels of non-invariance ranged from 0.00 to |± 0.033|. CFI differences smaller than |± 0.005| may be recommended when testing the metric invariance with ordinal data, and CFI differences smaller than |± 0.01| may be recommended when testing the scalar invariance with ordinal data.

Specifically, cutoff values examined in this study including |± 0.002|, |± 0.003| or |± 0.005| may be used by applied researchers when testing metric invariance (see Table 4.10), and -0.01 may be recommended when testing the scalar invariance (see Table 4.11). It should be noted that the performance of these cutoff values was not equal across studied conditions. In several conditions, these cutoff values failed, and thus, should not be relied upon by applied researchers. For example, even though using the smallest cutoff value |±0.002|, rejection rates were extremely low given both 8-indicator and 16-indicator models with sample size of 1200, 50% non-invariant items, and with symmetric

thresholds for lack of item loading invariance. In such circumstances, all proposed cutoff values are not recommended to be used.

4.2.2 RMSEA

Sample size. Descriptive results of RMSEA differences are shown in Tables 4.6 and 4.7, Figures 4.5, 4.6, 4.7 and 4.8. The sample size did not have a large impact on changes in RMSEA, although there was a slight increase in ΔRMSEA across several conditions. However, similar to CFI, the sample size impacted standard deviation of changes in RMSEA significantly. For example, when testing factor threshold invariance, given an 8-indicator model with symmetric thresholds and 25% non-invariant item loadings, the means of RMSEA changes varied from 0.009 to -0.080 between sample size 200 and sample size 1200. However, the standard deviation decreased from 0.031 with a sample size of 200 to 0.005 with a sample size of 1200.

Number of indicators. The number of indicators did not have an appreciable impact on change in the RMSEA index across successive tests. The results of ΔRMSEA were inconsistent across study conditions in the context of both lack of metric invariance and scalar invariance. For example, changes in RMSEA were larger in the 8-indicator models than in 16-indicator models with sample size of 200 when testing item threshold invariance, while the pattern was the opposite when factor loading invariance was examined. The standard deviations were higher in the 8-indicator conditions than the 16-indicator conditions.

Source of non-invariance. When the location of non-invariant items is on item loadings only, as expected, the changes in RMSEA were bigger when testing for factor loading non-invariance than testing factor threshold non-invariance. In contrast, when the

location of non-invariant items is on item thresholds only, the changes in RMSEA were smaller when testing factor loading non-invariance than when testing factor threshold non-invariance. Also, the results indicated that ΔRMSEAs were larger and more sensitive to test factor thresholds than factor loadings across the study conditions. For example, given the 8-indicator model with symmetric threshold, and sample size 200, the means of ΔRMSEA were 0.026 and 0.002 when testing factor loading invariance (vs.0.009 and 0.059 when testing factor threshold invariance) with 25% non-invariant item loadings and thresholds.

Levels of threshold symmetry. Overall, the changes in RMSEA were larger in the symmetric threshold model than the extreme asymmetric threshold model for both lack of factor loading invariance and lack of factor threshold invariance. This result indicated that ΔRMSEA is more sensitive to test symmetric threshold non-invariance than extreme asymmetric threshold non-invariance. For example, given a 16-indicator model with a sample size of 200 and 25% non-invariant item loadings, the means of ΔRMSEA were 0.030 for symmetric threshold vs. 0.021 for extreme asymmetric threshold conditions when testing factor loading invariance. When testing factor threshold invariance, the means of ΔRMSEA were 0.036 for symmetric threshold condition vs. 0.022 for extreme asymmetric threshold condition given a 16-indicator model with a sample size of 200 and 25% non-invariant item loadings.

Proportion of non-invariant items. In regard to the proportion of non-invariant item loadings, similar with CFI, changes in RMSEA were bigger with the 25% non-invariant item loading condition than when 50% of non-invariant items were present. However, the pattern of changes in RMSEA was opposite under a lack of threshold

invariance. For example, given a 16-indicator model with sample size of 1200 and extreme asymmetric data, the change in RMSEA were 0.033 when the non-invariant item loadings were 25% (vs. <0.001 when the non-invariant item loadings were 50%). However, the changes in RMSEA were -0.001 when the non-invariant item loadings were 25% vs. 0.014 when the non-invariant item loadings were 50% for lack of threshold invariance under the same conditions.

In terms of proportion of non-invariant item thresholds, changes in RMSEA were small and consistent across the 32 study conditions when testing item loading invariance. This was not unexpected as non-invariant item thresholds were not simulated in the population models. The change in RMSEA were bigger with 25% non-invariant item threshold conditions than with 50% non-invariant item threshold conditions, except for the 16-indicator models with extreme asymmetric thresholds. For example, given the 16-indicator model with symmetric item thresholds and sample size 1200, the changes in RMSEA were 0.054 when the non-invariant item thresholds were 25% vs. 0.020 when the non-invariant item thresholds were 50%. However, given the same model but with extreme asymmetric item thresholds, the changes in RMSEA were 0.038 with the 25% non-invariant item thresholds vs. 0.053 with the 50% non-invariant item thresholds.

Rejection Rates. Rejection rates for changes in RMSEA across sample sizes and levels of invariance are shown in Tables 4.10 and 4.11. Based on literature review and results of Study 1, six cutoff values were examined: 1) |±0.001| from Study 1; 2) 0.007 (Mead et al, 2008); 3) 0.01 (Chen, 2007); 4) 0.015 (Chen, 2007); 5) |± 0.02| from Study 1; 6) 0.05 (Rutkowski & Svetina, 2017).

First, considering the impact of sample size, the rejection rates of ΔRMSEA are consistent with ΔCFI, indicating that there was a large variation across study conditions with different sample sizes. Specifically, rejection rates of ΔRMSEA tended to increase in many study conditions when sample size increased for testing factor threshold invariance. However, rejection rates of ΔRMSEA decreased dramatically for sample size 1200 when factor loading invariance was examined. For example, given the 16-indicator model with symmetric and 50% non-invariant item loadings, the rejection rates were 45.9% when the sample size was 200 (vs. 19.8% when the sample size was 1200) using 0.007 as the cutoff value.

Second, the number of indicators did not impact rejection rates across study conditions. For example, given a model with sample size of 1200, symmetric and 25% non-invariant item thresholds for lack of threshold invariance, the rejection rates in RMSEA were 100.0% for the 8-indicator model vs. 100.0% for the 16-indicator model using a cutoff value of 0.007.

Third, concerning the source of non-invariance, similar to CFI, the rejection rates were higher when the source of non-invariance was from item loadings than item thresholds for lack of loading invariance, whereas the rejections rates were inconsistent when the source of non-invariance was from item thresholds for lack of threshold invariance. For example, given an 8-indicator model with symmetric thresholds, and sample size of 200, the rejection rates in CFI were 83.2% with 25% of non-invariant item loadings vs. 62.5% with 25% of non-invariant item thresholds when testing factor loading non-invariance using 0.007 as a cutoff value. The rejection rates, which performed inconsistently, were 76.8% and 92.9% with 25 % and 50% of non-invariant item loadings

vs. 99.5% and 62.2% with 25 % and 50% of non-invariant item thresholds when testing factor threshold non-invariance using 0.007 as a cutoff value.

Fourth, the impact of levels of threshold symmetry on the rejection rates for ΔRMSEA was not appreciable. For example, given a 16-indicator model with sample size of 200 and 25% non-invariant item loadings, the rejection rates in ΔRMSEA were 94.6% vs. 86.8% for symmetric and extreme asymmetric threshold conditions when testing factor loading non-invariance using 0.007 as a cutoff value. However, when sample size increased to 1200, rejection rates were 100% vs. 100% using 0.007 as the cutoff value.

Last, considering the proportion of non-invariant items with both item loadings and item thresholds, 25% of non-invariant items had higher rejection rates than 50% non-invariant items in many study conditions. For example, given a 16-indicator model with sample size of 1200 and extreme asymmetric threshold when testing lack of loading invariance, the rejection rates in RMSEA were 100.0% when there were 25% non-invariant item loadings vs. 17.9% when the non-invariant item loadings were 50% using 0.007 as a cutoff value.

Overall, on average, the RMSEA was more sensitive to detect threshold non-invariance (scalar non-invariance) than loading non-invariance (metric non-invariance), especially when the sample size was 1200. The average RMSEA difference across the conditions and across two levels of non-invariance ranged from |-0.001| to 0.087. Therefore, RMSEA differences smaller than |± 0.01| may be recommended when testing metric invariance with ordinal data, and RMSEA differences smaller than |± 0.02| may be recommended when testing scalar invariance.

Specifically, cutoff values examined in this study including |±0.001|, |± 0.007|, or |± 0.01| may be used by applied researchers based on different model conditions when testing metric invariance (see Table 4.10), and |±0.015| or |±0.02| may be suggested when testing scalar invariance (see Table 4.11). Similar to CFI, the proposed cutoff values in ΔRMSEA did not perform equally well across studied conditions. The cutoff values might fail in some conditions, and therefore should not be used. For example, using cutoff value of 0.007 recommended by Mead et al (2008), rejection rates were low given both 8-indicator and 16-indicator models with both sample size 200 and 1200, 50% non-invariant items, and with both symmetric and extreme asymmetric thresholds for lack of item loading invariance. In such circumstances, this study suggested |±0.001| as the criterion.

4.2.3 SRMR

Sample size. Descriptive results of SRMR differences are shown in Tables 4.8 and 4.9, Figures 4.5, 4.6, 4.7 and 4.8. The impact of sample size on ΔSRMRs was not appreciable. However, standard deviation of changes in SRMR decreased substantially when the sample size increased from 200 to 1200.

For example, given the 8-indicator model with symmetric thresholds and 25% non-invariant item loadings, the means of changes in SRMR were 0.012 and 0.014 for sample size 200 and sample size of 1200 when testing factor loading invariance. However, the means of changes in SRMR were 0.008 and 0.007 when testing item threshold invariance. Standard deviations decreased from 0.009 to 0.003 compared to samples of 200 and 1200 with the same study condition when testing factor loading invariance.

Number of indicators. The number of indicators did not have a large impact on the change in SRMR under factor loading invariance and threshold invariance. For example, given a model with symmetric thresholds, 25% non-invariant item loadings and sample size 200, the means of SRMR changes were 0.008 with the 8-indicator model (vs. 0.004 with the 16-indicator model) for factor threshold invariance testing, whereas the means of ΔSRMR were 0.012 vs. 0.014 for factor loading invariance testing. The standard deviations were higher in the 8-indicator conditions than the 16-indicator conditions.

Source of non-invariance. When non-invariant items are located on item loadings only, as expected, the changes in SRMR were bigger than non-invariant items on thresholds when testing factor loading non-invariance than testing factor threshold non-invariance. For example, given the 16-indicator model with sample size 200 and extreme asymmetric loadings, the change in SRMR was 0.015 for 25% non-invariant loadings vs. 0.007 for 25% non-invariant thresholds. In contrast, when the non-invariant items were on thresholds only, the changes in SRMR were bigger than non-invariant items on loadings when testing threshold non-invariance than testing loading non-invariance. However, unlike ΔCFI and ΔRMSEA, ΔSRMR was equally sensitive to tests of factor loading and factor threshold invariance across the study conditions.

Levels of threshold symmetry. The levels of threshold symmetry did not have an appreciable impact on changes in SRMR for both factor loading and threshold invariance testing. For example, given a 16-indicator model with sample size of 200 and 25% non-invariant item loadings, the means of SRMR changes were 0.014 for symmetric threshold condition and 0.015 for extreme asymmetric threshold condition when testing factor

loading invariance. However, the means of SRMR changes were 0.004 for symmetric threshold condition compared to 0.002 for extreme asymmetric condition.

Proportion of non-invariant items. Considering the proportion of non-invariant item loadings, changes in SRMR were bigger with 25% non-invariant item loading condition than with 50% non-invariant item loading condition when testing for lack of loading invariance, whereas the pattern was inconsistent for lack of threshold invariance. For example, given the 16-indicator model with sample size of 1200 and extreme asymmetric threshold, the change in SRMR averaged 0.016 when 25% of the items were non-invariant (vs. 0.003 when the non-invariant item loadings were 50%) for lack of loading invariance; however, the changes in SRMR were 0.001 vs. 0.002 for lack of threshold invariance with the same simulated condition. In terms of the proportion of non-invariant item thresholds, changes in SRMR were small and consistent across the 32 study conditions when testing item loading invariance. The pattern of changes in SRMR was inconsistent across the different proportion conditions tested.

Rejection Rates. Rejection rates for changes in SRMR across sample sizes and levels of invariance are shown in Tables 4.10 and 4.11. Based on literature review and results of Study 1, four cutoff values were examined: 1) 0.002 from Study 1; 2) 0.007 from Study 1; 3) 0.01 (Chen, 2007); and 4) 0.025 (Chen, 2007).

First, considering the impact of sample size, the results indicated that rejection rates based on ΔSRMR appeared to vary across study conditions with different sample sizes and with different cutoff values. The rejection rates of ΔSRMR decreased substantially for sample size 1200 when factor loading invariance was examined.

Second, for the number of indicators, the rejection rates did not change significantly between 8-indicator models and 16-indicator models. For example, given a model with symmetric item loadings, sample size of 200 for lack of loading invariance, and 25% non-invariant item loadings, the rejection rates in SRMR were 93.6% for the 8-indicator model vs. 100.0% for the 16-indicator model using a cutoff value of 0.002.

Third, the rejection rates were higher and more consistent across study conditions when the source of non-invariance was from item thresholds than item loadings, and when sample size was 200 than 1200. For example, given an 8-indicator model with symmetric thresholds, and sample size of 1200 testing loading invariance, the rejection rates in SRMR were 100.0% with 25 % of non-invariant item loadings vs. 50.2% with 25 % of non-invariant item thresholds when testing factor loading non-invariance using 0.002 as a cutoff value. However, given an 8-indicator model with symmetric threshold, and sample size of 200 for lack of threshold invariance, the rejection rates in SRMR were 91.7% with 25 % of non-invariant item loadings vs. 100.0% with 25 % of non-invariant item thresholds when testing factor threshold non-invariance using 0.002 as a cutoff value.

Fourth, the levels of threshold symmetry were not an impactful factor on the rejection rates for ΔSRMR. For example, given a 16-indicator model with sample size of 200 and 50% non-invariant item loadings for lack of loading invariance, the rejection rates in SRMR change were 99.3% vs. 99.8% for symmetric and extreme asymmetric threshold conditions when testing factor loading non-invariance using 0.002 as a cutoff value.

Last, regarding the proportion of non-invariant items with both item loadings and item thresholds, 25% of non-invariant items had higher rejection rates than 50% non-invariant items in the majority of study conditions. For example, given a 16-indicator model with sample size of 1200 and extreme asymmetric threshold for lack of item loading invariance, the rejection rates in SRMR were 100.0% when there were 25% non-invariant item loadings vs. 85.0% when the non-invariant item loadings were 50% using 0.002 as a cutoff value.

Overall, on average, the SRMR was more sensitive to identify threshold non-invariance (scalar non-invariance) in most study conditions than loading non-invariance (metric non-invariance). The average SRMR difference across the conditions and two levels of non-invariance ranged from 0.001 to 0.018. After examining the above cutoff values of SRMR differences, 0.002 or 0.007 are recommended when testing the metric invariance and scalar invariance given the high rejection rates in most studied conditions for both lack of loading invariance and threshold invariance. However, it is worth noting that applied researchers may use different cutoff values based on different model conditions. For example, when testing lack of loading invariance, given both 8-indicator and 16-indicator models with both sample size 200 and 1200, 50% non-invariant items, and with both symmetric and extreme asymmetric thresholds, cutoff value of 0.002 is recommended to use than cutoff value of 0.007.

Chapter 5 presented the discussion of findings, conclusions, and implications based on the results. Summaries were conducted to compare results with previous research on changes in the three fit indices within the framework of multiple group CFA invariance testing. Implications and recommendations of cutoff values on the model fit

changes across study conditions were discussed in this context. Last, limitations and future research were included in this chapter.

Table 4.3 Convergence rates across study conditions

# Indicators, threshold symmetry, % non-invariance			Convergence Rate (%)					
			Sample size 200			Sample size 1200		
			Configural	Metric	Scalar	Configural	Metric	Scalar
8 Indicators	Symmetry	25% loadings	99.6	99.8	100.0	100.0	100.0	100.0
		25% thresholds	99.8	100.0	100.0	100.0	100.0	100.0
		50% loadings	99.9	100.0	100.0	100.0	100.0	100.0
		50% thresholds	97.1	97.2	97.2	100.0	100.0	100.0
	Extreme Asymmetry	25% loadings	98.7	99.8	100.0	100.0	100.0	100.0
		25% thresholds	98.5	98.5	98.5	100.0	100.0	100.0
		50% loadings	99.6	100.0	100.0	100.0	100.0	100.0
		50% thresholds	97.0	97.2	97.2	100.0	100.0	100.0
16 Indicators	Symmetry	25% loadings	100.0	100.0	100.0	100.0	100.0	100.0
		25% thresholds	97.5	97.5	97.5	100.0	100.0	100.0
		50% loadings	100.0	100.0	100.0	100.0	100.0	100.0
		50% thresholds	**94.6**	**94.6**	**94.6**	100.0	100.0	100.0
	Extreme Asymmetry	25% loadings	100.0	100.0	100.0	100.0	100.0	100.0
		25% thresholds	97.5	97.5	97.5	100.0	100.0	100.0
		50% loadings	100.0	100.0	100.0	100.0	100.0	100.0
		50% thresholds	**94.6**	**94.6**	**94.6**	100.0	100.0	100.0

Note. The lowest convergence rates indicated by bold text.

Table 4.4 Mean, standard deviation and the 1st and 5th percentiles of CFI difference for testing loading invariance by study conditions

# Indicators, threshold symmetry, % non-invariance			Sample size 200				Sample size 1200			
			M	SD	5th	1st	M	SD	5th	1st
8 Indicators	Symmetry	25% loadings	-0.007	0.010	-0.014	-0.020	-0.010	0.003	-0.014	-0.016
		25% thresholds	-0.001	0.006	-0.002	-0.004	0.000	0.000	0.000	0.000
		50% loadings	-0.001	0.006	-0.002	-0.003	0.000	0.001	0.000	0.000
		50% thresholds	-0.001	0.004	-0.003	-0.001	0.000	0.000	0.000	0.000
	Extreme Asymmetry	25% loadings	-0.007	**0.014**	-0.014	**-0.021**	-0.010	**0.004**	-0.015	-0.017
		25% thresholds	-0.001	0.004	-0.001	-0.002	0.000	0.001	0.000	0.000
		50% loadings	-0.001	0.012	-0.003	-0.006	0.000	0.001	0.000	0.000
		50% thresholds	-0.001	0.007	-0.003	-0.002	0.000	0.001	0.000	-0.001
16 Indicators	Symmetry	25% loadings	**-0.011**	0.006	-0.016	-0.021	**-0.014**	0.003	**-0.017**	-0.018
		25% thresholds	-0.001	0.002	-0.002	-0.002	0.000	0.000	0.000	0.000
		50% loadings	-0.001	0.003	-0.002	-0.003	0.000	0.001	0.000	0.000
		50% thresholds	-0.001	0.002	-0.001	-0.001	0.000	0.000	0.000	0.000
	Extreme Asymmetry	25% loadings	**-0.011**	0.007	**-0.017**	-0.020	-0.013	0.003	**-0.017**	**-0.019**
		25% thresholds	-0.001	0.003	-0.002	0.000	0.000	0.001	0.000	0.000
		50% loadings	-0.001	0.004	-0.001	-0.003	0.000	0.001	0.000	0.000
		50% thresholds	-0.001	0.003	-0.002	0.000	0.000	0.001	0.000	0.000

Note. Largest difference in each column indicated by bold text.

Table 4.5 Mean, standard deviation and the 1st and 5th percentiles of CFI difference for testing loading and threshold invariance by study conditions

# Indicators, threshold symmetry, % non-invariance			Sample size 200				Sample size 1200			
			M	SD	5th	1st	M	SD	5th	1st
8 Indicators	Symmetry	25% loadings	-0.008	**0.013**	-0.013	-0.018	-0.010	0.003	-0.012	-0.013
		25% thresholds	**-0.033**	**0.013**	**-0.048**	**-0.058**	**-0.044**	**0.006**	**-0.052**	**-0.054**
		50% loadings	-0.017	0.009	-0.028	-0.030	-0.022	0.004	-0.027	-0.028
		50% thresholds	-0.004	0.005	-0.008	-0.009	-0.006	0.002	-0.009	-0.010
	Extreme Asymmetry	25% loadings	-0.003	**0.013**	-0.007	-0.009	-0.004	0.002	-0.005	-0.005
		25% thresholds	-0.021	0.010	-0.030	-0.035	-0.030	0.005	-0.036	-0.039
		50% loadings	-0.006	0.007	-0.012	-0.016	-0.014	0.004	-0.018	-0.019
		50% thresholds	-0.032	**0.013**	-0.046	-0.051	-0.017	0.003	-0.021	-0.022
16 Indicators	Symmetry	25% loadings	-0.005	0.003	-0.006	-0.008	-0.007	0.001	-0.007	-0.008
		25% thresholds	-0.016	0.005	-0.019	-0.024	-0.024	0.003	-0.028	-0.029
		50% loadings	-0.010	0.004	-0.013	-0.015	-0.014	0.002	-0.016	-0.017
		50% thresholds	-0.002	0.002	-0.004	-0.005	-0.004	0.001	-0.005	-0.006
	Extreme Asymmetry	25% loadings	-0.002	0.002	-0.002	-0.002	-0.002	0.001	-0.003	-0.003
		25% thresholds	-0.011	0.004	-0.014	-0.016	-0.018	0.002	-0.021	-0.022
		50% loadings	-0.003	0.003	-0.004	-0.006	-0.004	0.001	-0.005	-0.005
		50% thresholds	-0.019	0.006	-0.024	-0.025	-0.031	0.003	-0.035	-0.037

Note. Largest difference in each column indicated by bold text.

Table 4.6 Mean, standard deviation and the 1st and 5th percentiles of RMSEA difference for testing loading invariance by study conditions

# Indicators, threshold symmetry, % non-invariance			Sample size 200				Sample size 1200			
			M	SD	95th	99th	M	SD	95th	99th
8 Indicators	Symmetry	25% loadings	0.026	**0.036**	0.022	0.022	**0.044**	**0.011**	**0.040**	**0.038**
		25% thresholds	0.002	0.029	0.000	0.000	0.000	0.007	0.000	-0.001
		50% loadings	0.002	0.025	-0.001	-0.001	-0.001	0.007	-0.001	-0.002
		50% thresholds	0.002	0.024	0.002	-0.001	0.000	0.006	-0.002	-0.001
	Extreme Asymmetry	25% loadings	0.019	**0.036**	0.015	0.020	0.035	**0.011**	0.032	0.032
		25% thresholds	0.001	0.016	0.000	0.000	0.000	0.006	0.000	-0.002
		50% loadings	0.001	0.033	-0.001	0.004	0.000	0.007	-0.002	-0.001
		50% thresholds	0.002	0.029	0.001	0.001	0.000	0.006	0.000	0.000
16 Indicators	Symmetry	25% loadings	**0.030**	0.015	**0.024**	**0.026**	0.041	0.007	0.038	0.037
		25% thresholds	0.003	0.010	0.002	0.003	0.000	0.004	-0.001	-0.001
		50% loadings	0.002	0.011	0.002	0.001	0.000	0.005	0.000	0.000
		50% thresholds	0.003	0.010	0.002	0.003	0.000	0.004	0.000	0.000
	Extreme Asymmetry	25% loadings	0.021	0.013	0.018	0.018	0.033	0.007	0.029	0.028
		25% thresholds	0.002	0.010	0.002	0.001	0.000	0.005	0.000	0.000
		50% loadings	0.001	0.011	0.001	0.001	0.000	0.005	0.000	-0.001
		50% thresholds	0.002	0.010	0.002	-0.001	0.000	0.004	0.000	0.001

Note. Largest difference in each column indicated by bold text.

Table 4.7 Mean, standard deviation and the 1st and 5th percentiles of RMSEA difference for testing loading and threshold invariance by study conditions

# Indicators, threshold symmetry, % non-invariance			Sample size 200				Sample size 1200			
			M	SD	95th	99th	M	SD	95th	99th
8 Indicators	Symmetry	25% loadings	0.009	**0.031**	0.000	-0.003	0.008	0.005	0.006	0.006
		25% thresholds	**0.059**	0.022	**0.040**	**0.031**	**0.087**	0.009	**0.077**	**0.076**
		50% loadings	0.037	0.020	0.019	0.016	0.054	0.009	0.045	0.044
		50% thresholds	0.012	0.018	0.002	-0.005	0.028	0.009	0.022	0.021
	Extreme Asymmetry	25% loadings	-0.001	0.028	-0.008	-0.013	-0.001	0.004	-0.005	-0.005
		25% thresholds	0.040	0.019	0.021	0.015	0.060	**0.010**	0.049	0.046
		50% loadings	0.008	0.016	0.000	-0.002	0.033	0.008	0.024	0.022
		50% thresholds	0.054	0.020	0.032	0.027	0.044	0.009	0.033	0.030
16 Indicators	Symmetry	25% loadings	0.003	0.005	0.001	-0.001	0.005	0.002	0.004	0.004
		25% thresholds	0.036	0.011	0.023	0.020	0.054	0.006	0.047	0.045
		50% loadings	0.023	0.010	0.013	0.012	0.037	0.005	0.030	0.029
		50% thresholds	0.007	0.008	0.003	0.001	0.020	0.005	0.014	0.013
	Extreme Asymmetry	25% loadings	-0.001	0.004	-0.003	-0.006	-0.001	0.001	-0.001	-0.001
		25% thresholds	0.022	0.010	0.012	0.009	0.038	0.005	0.031	0.029
		50% loadings	0.004	0.007	0.000	0.000	0.014	0.004	0.008	0.008
		50% thresholds	0.033	0.011	0.020	0.019	0.053	0.006	0.045	0.043

Note. Largest difference in each column indicated by bold text.

Table 4.8 Mean, standard deviation and the 1st and 5th percentiles of SRMR difference for testing loading invariance by study conditions

# Indicators, threshold symmetry, % non-invariance			Sample size 200				Sample size 1200			
			M	SD	95th	99th	M	SD	95th	99th
8 Indicators	Symmetry	25% loadings	0.012	0.009	0.015	0.016	0.014	0.003	0.016	0.017
		25% thresholds	0.006	0.007	0.008	0.010	0.002	0.001	0.003	0.003
		50% loadings	0.007	0.006	0.008	0.009	0.003	0.001	0.003	0.003
		50% thresholds	0.006	0.005	0.007	0.007	0.002	0.001	0.003	0.003
	Extreme Asymmetry	25% loadings	0.012	**0.011**	0.015	0.017	0.013	**0.004**	0.016	0.017
		25% thresholds	0.007	0.004	0.007	0.007	0.003	0.001	0.003	0.003
		50% loadings	0.008	**0.011**	0.011	0.010	0.003	0.002	0.004	0.004
		50% thresholds	0.007	0.009	0.008	0.009	0.003	0.001	0.003	0.004
16 Indicators	Symmetry	25% loadings	0.014	0.004	**0.017**	**0.018**	**0.018**	0.003	**0.020**	**0.020**
		25% thresholds	0.006	0.002	0.006	0.007	0.002	0.001	0.003	0.003
		50% loadings	0.007	0.003	0.008	0.009	0.003	0.001	0.004	0.004
		50% thresholds	0.006	0.002	0.006	0.007	0.002	0.001	0.003	0.003
	Extreme Asymmetry	25% loadings	**0.015**	0.005	**0.017**	**0.018**	0.016	0.003	0.019	0.019
		25% thresholds	0.007	0.003	0.008	0.009	0.003	0.001	0.003	0.004
		50% loadings	0.009	0.003	0.009	0.012	0.003	0.001	0.004	0.004
		50% thresholds	0.007	0.003	0.008	0.007	0.003	0.001	0.003	0.004

Note. Largest difference in each column indicated by bold text.

Table 4.9 Mean, standard deviation and the 1st and 5th percentiles of SRMR difference for testing loading and threshold invariance by study conditions

# Indicators, threshold symmetry, % non-invariance			Sample size 200				Sample size 1200			
			M	SD	95th	99th	M	SD	95th	99th
8 Indicators	Symmetry	25% loadings	0.008	**0.009**	0.008	0.008	0.007	**0.002**	0.008	0.008
		25% thresholds	0.009	0.002	0.008	0.007	0.013	**0.002**	0.013	0.013
		50% loadings	0.011	0.003	0.010	0.010	0.014	**0.002**	0.013	0.013
		50% thresholds	0.006	0.002	0.005	0.006	0.006	0.001	0.005	0.005
	Extreme Asymmetry	25% loadings	0.004	**0.009**	0.003	0.005	0.003	0.001	0.002	0.003
		25% thresholds	0.010	0.003	0.010	0.009	**0.016**	**0.002**	**0.015**	**0.014**
		50% loadings	0.004	0.002	0.004	0.003	0.007	0.001	0.006	0.006
		50% thresholds	**0.016**	0.004	**0.016**	**0.014**	0.011	**0.002**	0.013	0.009
16 Indicators	Symmetry	25% loadings	0.004	0.001	0.004	0.005	0.004	0.001	0.004	0.004
		25% thresholds	0.005	0.001	0.005	0.005	0.008	0.001	0.007	0.007
		50% loadings	0.005	0.001	0.006	0.005	0.008	0.001	0.007	0.007
		50% thresholds	0.003	0.001	0.003	0.003	0.003	0.001	0.003	0.002
	Extreme Asymmetry	25% loadings	0.002	0.001	0.002	0.002	0.001	0.000	0.001	0.001
		25% thresholds	0.006	0.001	0.005	0.004	0.010	0.001	0.009	0.009
		50% loadings	0.002	0.001	0.002	0.002	0.002	0.000	0.002	0.002
		50% thresholds	0.009	0.002	0.008	0.008	**0.016**	0.001	**0.015**	**0.014**

Note. Largest difference in each column indicated by bold text

Table 4.10 Rejection rates based on changes in fit indices between metric and configural models

# indicators	Symmetry levels	% non-invariant	Sample size 200											
			ΔCFI (%)			ΔRMSEA (%)				ΔSRMR (%)				
			|± 0.002|	|± 0.003|	-0.005	|±0.001|	0.007	0.01	0.015	0.05	0.002	0.007	0.01	0.025
8 Indicators	Symmetry	25% loadings	77.5	72.3	61.9	90.0	83.2	79.0	74.2	31.9	93.6	72.6	59.1	6.4
		25% thresholds	51.0	41.2	28.6	70.1	62.5	57.4	47.8	9.8	88.1	46.7	25.9	1.0
		50% loadings	47.0	39.4	26.0	66.2	56.1	51.5	42.8	7.6	91.8	44.9	25.0	1.2
		50% thresholds	39.0	29.5	16.8	63.6	52.7	47.4	40.1	6.2	88.1	35.4	16.7	0.4
	Extreme Asymmetry	25% loadings	78.1	73.5	64.1	89.5	80.6	77.1	70.7	26.0	92.4	70.7	57.4	12.3
		25% thresholds	38.6	28.8	14.8	58.0	46.5	39.2	29.0	0.5	91.9	38.3	17.3	0.0
		50% loadings	63.9	56.8	45.5	78.4	70.5	65.8	59.8	14.5	89.2	60.5	47.5	6.4
		50% thresholds	54.5	45.8	31.5	71.9	62.9	58.1	51.6	11.1	89.7	50.1	33.3	2.7
16 Indicators	Symmetry	25% loadings	95.2	92.1	82.9	98.5	94.6	91.5	84.8	10.0	100.0	96.5	85.3	1.7
		25% thresholds	27.3	14.4	4.4	65.3	37.0	27.2	15.7	0.0	98.6	29.8	5.3	0.0
		50% loadings	41.4	27.8	11.2	73.3	45.9	33.7	19.4	0.0	99.3	49.3	15.9	0.0
		50% thresholds	27.5	14.3	4.2	66.9	37.9	27.6	13.8	0.0	98.9	29.8	5.0	0.0
	Extreme Asymmetry	25% loadings	89.5	86.4	77.0	95.8	86.8	78.3	65.7	1.8	100.0	96.7	84.1	2.6
		25% thresholds	37.9	24.0	9.1	73.7	38.5	28.1	15.1	0.0	99.0	43.8	12.8	0.0
		50% loadings	51.7	38.6	19.7	76.9	43.1	29.8	14.9	0.0	99.8	63.6	31.2	1.0
		50% thresholds	35.4	22.0	8.5	71.0	37.8	26.2	13.9	0.0	99.0	43.5	12.5	0.0

# indicators	Symmetry levels	% non-invariant	Sample size 1200											
			ΔCFI			ΔRMSEA					ΔSRMR			
			|±0.002|	|±0.003|	-0.005	|±0.001|	0.007	0.01	0.015	0.05	0.002	0.007	0.01	0.025
8 Indicators	Symmetry	25% loadings	99.4	98.5	94.8	100.0	99.9	99.6	99.0	32.9	100.0	98.2	88.6	0.1
		25% thresholds	0.7	0.0	0.0	52.6	25.9	15.8	5.0	0.0	50.2	0.2	0.0	0.0
		50% loadings	1.8	0.2	0.0	54.3	30.7	19.9	6.1	0.0	63.2	1.1	0.0	0.0
		50% thresholds	0.6	0.0	0.0	53.7	25.0	12.9	3.7	0.0	51.3	0.1	0.0	0.0
	Extreme Asymmetry	25% loadings	98.8	96.8	88.8	99.6	99.2	98.6	96.0	7.9	100.0	96.2	78.5	0.1
		25% thresholds	2.2	0.4	0.0	51.1	23.7	13.3	4.9	0.0	59.5	0.7	0.0	0.0
		50% loadings	6.0	1.4	0.2	54.4	28.5	17.0	5.2	0.0	70.3	2.9	0.2	0.0
		50% thresholds	2.1	0.4	0.0	52.1	24.2	12.7	4.7	0.0	58.6	0.4	0.0	0.0
16 Indicators	Symmetry	25% loadings	100.0	100.0	100.0	100.0	100.0	100.0	100.0	9.5	100.0	100.0	99.9	0.5
		25% thresholds	0.1	0.0	0.0	53.0	13.8	5.1	0.3	0.0	61.0	0.0	0.0	0.0
		50% loadings	0.7	0.1	0.0	57.4	19.8	6.8	0.9	0.0	79.8	0.2	0.0	0.0
		50% thresholds	0.1	0.0	0.0	51.4	12.5	4.5	0.7	0.0	60.8	0.0	0.0	0.0
	Extreme Asymmetry	25% loadings	100.0	100.0	99.9	100.0	100.0	100.0	99.7	0.0	100.0	100.0	98.9	0.1
		25% thresholds	1.1	0.3	0.0	58.8	14.2	4.5	0.2	0.0	69.5	0.3	0.0	0.0
		50% loadings	2.7	0.7	0.0	59.0	17.9	7.4	0.8	0.0	85.0	1.6	0.0	0.0
		50% thresholds	0.9	0.1	0.0	55.7	13.6	3.8	0.3	0.0	69.0	0.1	0.0	0.0

Table 4.11 Rejection Rates based on changes in fit indices between scalar and metric models

# indictrs	Symtry levels	% non-invariant	Sample size 200													
			ΔCFI (%)					ΔRMSEA (%)				ΔSRMR (%)				
			|±0.002|	|±0.003|	-0.005	0.01	0.03	0.007	0.01	0.015	|±0.02|	0.05	0.002	0.007	0.01	0.025
8 Indictrs	Symtry	25% ldngs	86.4	82.4	71.2	47.3	5.3	76.8	67.3	54.8	46.1	13.2	91.7	58.0	37.3	3.3
		25% tshds	99.9	99.9	99.7	97.9	54.9	99.5	98.9	98.1	96.0	65.7	100.0	85.2	36.9	0.0
		50% ldngs	95.7	94.8	91.9	77.6	9.1	92.9	89.5	83.0	76.6	27.5	100.0	89.8	55.4	0.0
		50% tshds	58.9	50.6	36.7	14.3	0.0	62.2	54.4	42.8	32.6	2.9	99.5	27.0	3.8	0.0
	Extrm Asymtry	25% ldngs	79.2	70.3	56.5	34.0	4.5	74.8	63.7	51.1	38.9	11.5	89.1	32.5	21.9	2.8
		25% tshds	98.1	97.1	94.4	85.1	17.0	96.0	93.8	89.4	83.7	33.6	100.0	89.7	52.8	0.3
		50% ldngs	63.5	56.3	45.8	23.9	1.0	54.6	45.9	32.5	24.4	0.8	96.9	5.1	0.2	0.0
		50% tshds	100.0	100.0	100.0	98.2	51.5	99.4	99.2	97.6	95.7	56.9	100.0	99.6	93.8	2.0
16 Indictrs	Symtry	25% ldngs	96.2	74.4	42.2	2.9	0.0	21.0	8.1	2.4	0.9	0.0	99.2	1.8	4.6	0.0
		25% tshds	100.0	100.0	99.4	90.3	0.9	100.0	100.0	98.7	94.4	11.3	100.0	2.4	0.0	0.0
		50% ldngs	97.7	96.6	91.5	50.6	0.0	95.8	91.3	74.8	53.3	0.1	100.0	7.3	98.5	0.0
		50% tshds	52.8	35.1	13.5	0.4	0.0	45.9	33.8	17.9	8.7	0.0	91.3	0.0	0.0	0.0
	Extrm Asymtry	25% ldngs	43.6	25.6	8.2	0.1	0.0	7.5	2.2	1.1	0.4	0.0	35.0	0.0	0.0	0.0
		25% tshds	99.4	98.8	94.5	59.2	0.0	96.5	90.7	72.2	52.5	0.2	100.0	13.5	0.2	0.0
		50% ldngs	58.6	45.2	22.2	1.8	0.0	28.9	18.6	8.9	3.5	0.0	55.9	0.0	1.7	0.0
		50% tshds	100.0	100.0	99.9	96.5	3.1	100.0	99.7	96.4	87.4	7.0	100.0	88.8	23.7	0.0

# indicts	Symtry levels	% non-invariant	Sample size 1200													
			ΔCFI						ΔRMSEA				ΔSRMR			
			|±0.002|	|±0.003|	-0.005	0.01	0.03	0.007	0.01	0.015	|± 0.02|	0.05	0.002	0.007	0.01	0.025
8 Indicts	Symtry	25% ldngs	100.0	100.0	97.4	44.9	0.0	52.7	30.8	9.5	20.0	0.0	100.0	55.4	4.6	0.0
		25% tshds	100.0	100.0	100.0	100.0	99.9	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.4	0.0
		50% ldngs	100.0	100.0	100.0	99.9	2.1	100.0	100.0	100.0	100.0	66.1	100.0	100.0	98.5	0.0
		50% tshds	97.5	92.2	68.1	5.4	0.0	99.5	98.7	93.3	81.5	0.1	100.0	15.0	0.1	0.0
	Extrm Asymtry	25% ldngs	85.0	64.3	26.0	0.5	0.0	9.5	1.6	0.4	0.0	0.0	84.8	0.0	0.0	0.0
		25% tshds	100.0	100.0	100.0	100.0	51.9	100.0	100.0	100.0	100.0	81.3	100.0	100.0	99.9	0.1
		50% ldngs	100.0	100.0	99.8	87.6	0.0	100.0	99.8	98.7	93.5	0.8	100.0	44.8	1.7	0.0
		50% tshds	100.0	100.0	100.0	99.1	0.0	100.0	100.0	100.0	99.8	30.2	100.0	98.7	63.4	0.0
16 Indicts	Symtry	25% ldngs	100.0	100.0	90.8	0.7	0.0	11.4	0.2	0.0	0.0	0.0	100.0	0.0	0.0	0.0
		25% tshds	100.0	100.0	100.0	100.0	1.6	100.0	100.0	100.0	100.0	11.3	100.0	86.4	0.5	0.0
		50% ldngs	100.0	100.0	100.0	98.5	0.0	100.0	100.0	100.0	100.0	0.0	100.0	81.5	0.4	0.0
		50% tshds	97.3	84.9	24.6	0.0	0.0	99.8	97.2	80.7	52.2	0.0	97.0	0.0	0.0	0.0
	Extrm Asymtry	25% ldngs	61.7	15.3	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.6	0.0	0.0	0.0
		25% tshds	100.0	100.0	100.0	100.0	0.0	100.0	100.0	100.0	100.0	0.0	100.0	100.0	43.5	0.0
		50% ldngs	94.1	80.3	19.5	0.0	0.0	94.8	77.7	42.7	7.8	0.0	68.2	0.0	0.0	0.0
		50% tshds	100.0	100.0	100.0	100.0	60.0	100.0	100.0	100.0	100.0	66.9	100.0	100.0	100.0	0.0

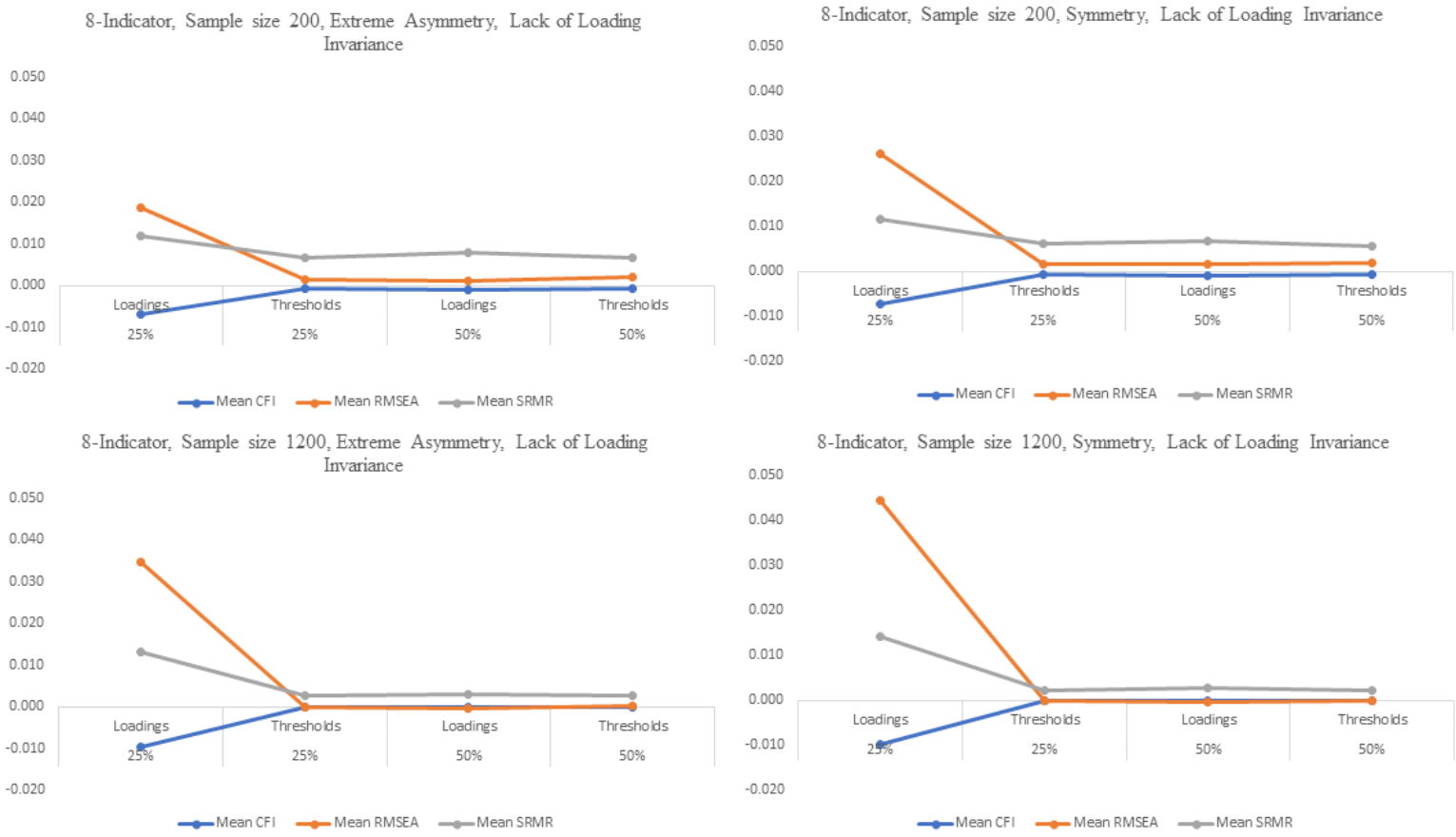

Figure 4.5 Mean changes in fit indices based on studied conditions of 8 indicators, sample size of 200 and 1200, symmetry and extreme symmetry, and across 25% and 50% of non-invariant loadings for factor loading non-invariance.

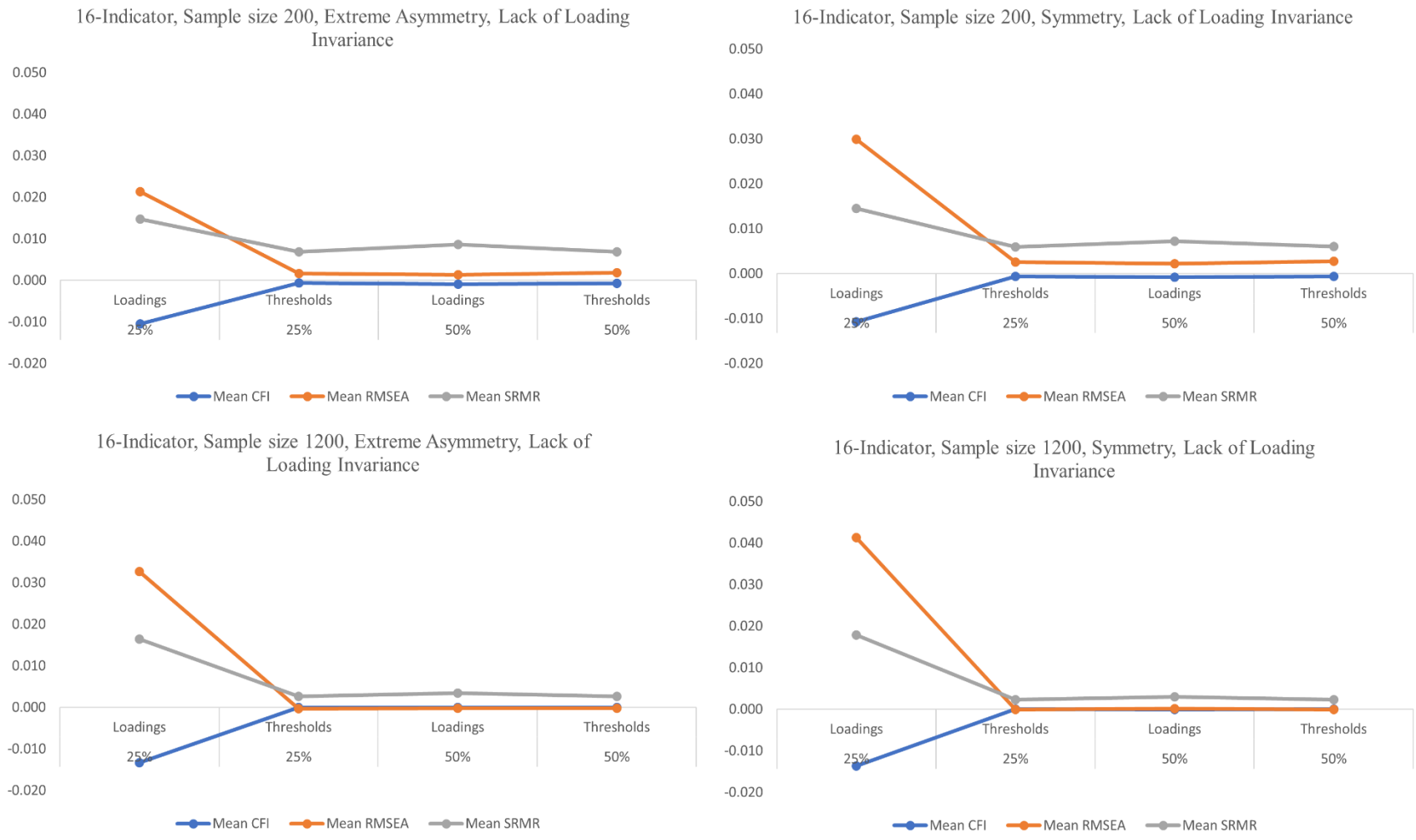

Figure 4.6 Mean changes in fit indices based on studied conditions of 16 indicators, sample size of 200 and 1200, symmetry and extreme symmetry, and across 25% and 50% of non-invariant loadings for factor loading non-invariance.

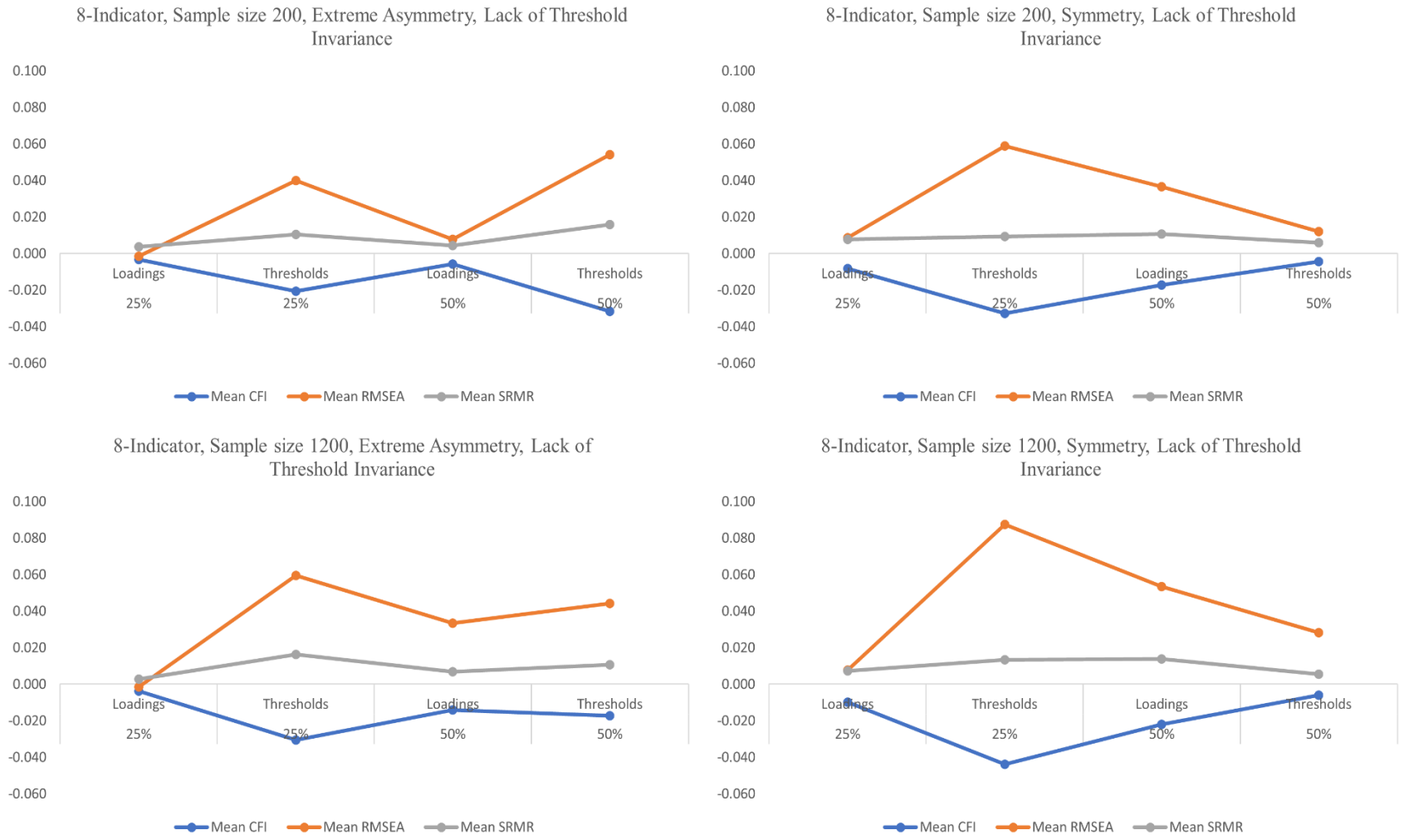

Figure 4.7 Mean changes in fit indices based on studied conditions of 16 indicators, sample size of 200 and 1200, symmetry and extreme symmetry, and across 25% and 50% of non-invariant loadings for factor threshold non-invariance

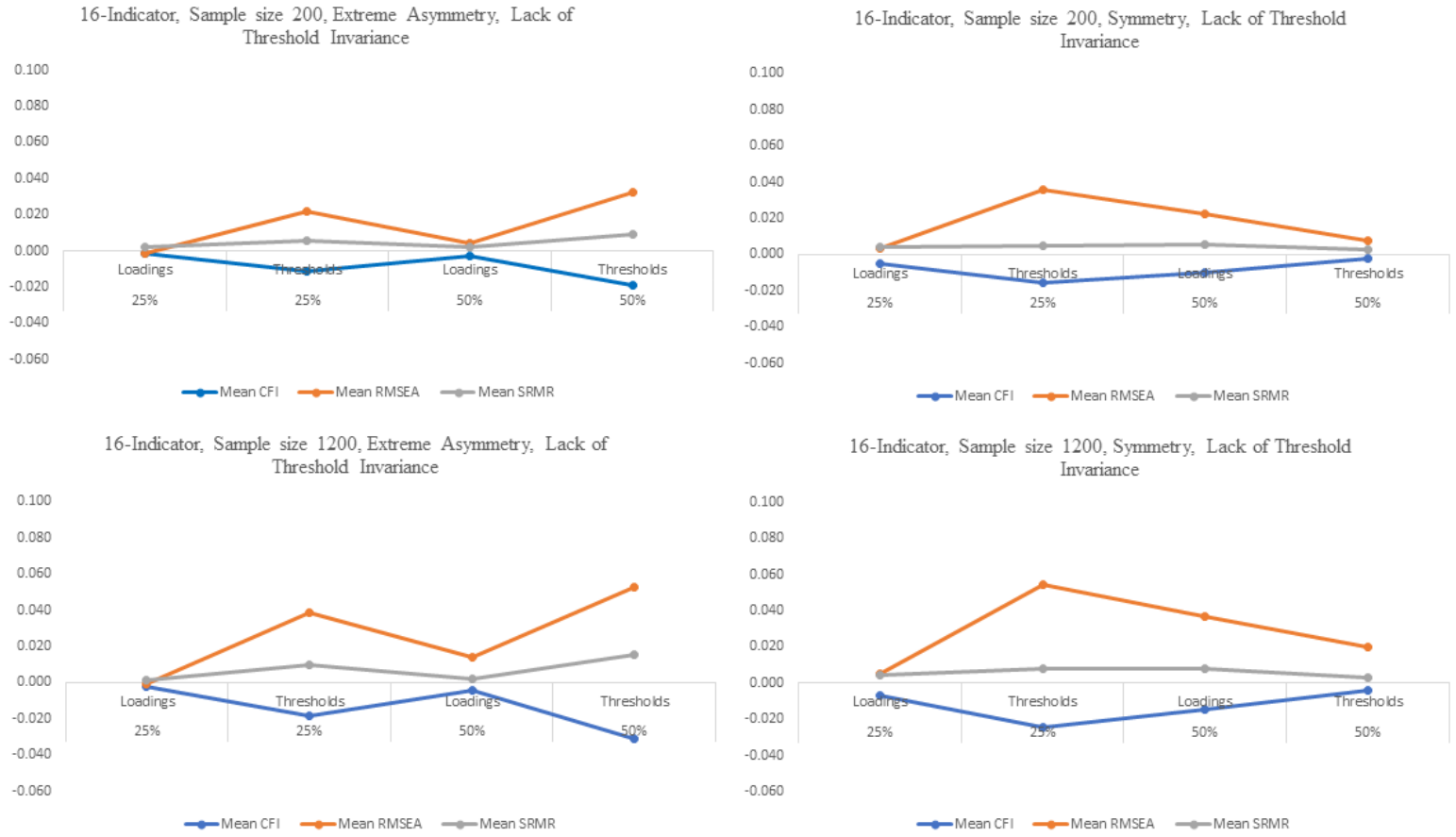

Figure 4.8 Mean changes in fit indices based on studied conditions of 16 indicators, sample size of 200 and 1200, symmetry and extreme symmetry, and across 25% and 50% of non-invariant loadings for factor threshold non-invariance

CHAPTER 5

DISCUSSION

This study examined the performance of three model fit indices commonly used with multiple-group CFA invariance testing. As many measurement instruments use ordinal data, the goal here was to examine performance of the indices when categorical data are analyzed. As increasing numbers of applied researchers are aware of the importance of measurement invariance prior to conducting group comparisons, it is critical to provide recommendations and guidelines when relying upon model fit indices to evaluate measurement invariance. However, very few studies have investigated this issue in the context of categorical-ordered data. To fill this gap, two Monte Carlo studies were conducted. Study 1 examined random variations of three model fit indices available in Mplus: CFI, RMSEA and SRMR under six levels of invariance including factor loadings, thresholds, residual variances, latent means, factor variances, and factor covariances. Based on Study 1 results, cutoff values were proposed for assisting researchers when using such indices to evaluate the presence of measurement invariance when testing a more constrained model and a less restricted model.

Study 2 examined the impact of five factors on the sensitivity of fit indices' changes to identify two levels of non-invariance which are commonly tested: metric non-invariance, and scalar non-invariance. In addition, rejection rates based on proposed and frequently used cutoff values in previous studies were tested in Study 2.

5.1 Performance of Model Fit Indices

First, in contrast to Chen's study (2007), which concluded that SRMR is more sensitive to random variation in factor loadings than in intercepts or residual variances, this study found SRMR to be equally sensitive to all three levels of invariance, especially when the sample size was at the lower level (i.e., 300). Therefore, unlike Chen's result (2007), which suggested two cutoff values for all three levels of invariance tests, only one cutoff value of SRMR was recommended for different levels of invariance tests. This inconsistency may be due to the two different data types and estimators examined across the two studies. Chen's study focused on assessing normally distributed continuous data with Maximum Likelihood estimation method. However, this study concentrated on examining non-normal categorical-ordered data using WLSMV estimator.

In addition, the findings from this study indicated that both CFI and RMSEA appear to be more sensitive to detecting non-invariance in thresholds than loading values. This result was not consistent with Chen's results (2007) and Cheung and Rensvold's results (2002), which concluded that CFI and RMSEA were equally sensitive to invariance in loadings, intercepts, and residual variances. However, it is noted that intercept values which are tested with continuous data are not equivalent to threshold values estimated when ordinal data are present. The finding of this study, however, was an echo of Sokolov' conclusions (2019), indicating that the CFI, RMSEA, and TLI are largely effective at identifying scalar non-invariance with categorical data using the WLSMV estimation method.

Overall, based on the above conclusions, different cutoff values were recommended for use with CFI and RMSEA across three invariance levels, but the same

cutoff value was recommended for SRMR across three levels of invariance tests. Specifically, when testing loading invariance and looking to support equality of groups, a change of ≤ |±0.003| is proposed for CFI, and a change of ≤ |±0.001| is recommended for RMSEA; When testing threshold and residual variance invariance levels, a change of ≤ |±0.03| is recommended for CFI, and a change of ≤ |±0.02| can be used for RMSEA. With respect to SRMR, a change of ≤ 0.007 is proposed across invariance levels of loadings, thresholds and residual variances.

Second, the finding of this study on the effect of group-level sample size is consistent with previous studies (Chen; 2007; Mead et al., 2008; Sokolov, 2019), indicating that the group-level sample size only slightly impacts the changes in all three model fit indices, and the impact is highly inconsistent across studied conditions. However, the group-level sample size substantially impacts standard deviations of the difference (i.e., changes in fit) for the three model fit indices across studied conditions. These results illustrated that as sample size increases, model estimation becomes more precise. As a result, changes in fit indices become small as sample size increases (Cheung & Rensvold, 2002). For this reason, a more conservative cutoff value (e.g., 0.002 in ΔCFI) may be used when the sample size is small (e.g., < 300) and a more liberal value (e.g., 0.005 in ΔCFI) may be reported when sample size is large (e.g., > 1,000). Additionally, the findings are similar to Chen's (2007) study, which concluded that the means, standard deviations, and percentiles of SRMR were larger in small samples as compared to the values produced by CFI and RMSEA. Last, among the three fit indices, ΔSRMR was the least sensitive to sample size. This finding is consistent with previous results that SRMR was relatively independent of sample size (e.g., Chen, 2007).

In terms of rejection rates, the current study found that changes in the three model fit indices appear to vary across studied conditions. Generally, as sample size increased, rejection rates based on fit indices tended to increase in most studied conditions. An exception was noted for conditions with sample size of 1200 and with 50% non-invariant item loadings when testing loading non-invariance. The low rejection rates may be caused by multiple factors manipulated in this study, such as proportion of non-invariant items, source of non-invariance, levels of threshold symmetry or the interactive effect among these factors.

Third, Chen (2007) found that changes in fit statistics were influenced by an interaction between the proportion of invariance and the pattern of invariance (whether lack of invariance was uniform or mixed). Specifically, when non-invariance was uniform, the relation between the proportion of invariance and changes in fit indices was non-monotonic. For example, when 0% and 75% of the items were invariant, the changes in CFI, RMSEA and SRMR were small, whereas when 50% of the items were invariant, the average change in fit indices was largest for testing lack of loading invariance. In contrast, when lack of loading invariance was mixed, the change in fit statistics was monotonic. For example, the results of this study noted that changes in the respective fit indices were larger when 50% of the items were invariant than when 75% of the items were invariant. In the present study, both changes in fit indices and rejection rates decreased as the proportion of non-invariant items increased from 25% (75% invariant items) to 50% (50% invariant items). In other words, the changes in fit statistics were bigger with 25% (75% invariant items) non-invariant items than 50% non-invariant items. These findings are opposite to Chen's findings for lack of loading non-invariance.

Regarding the lack of threshold invariance, the performance of fit indices was mixed across conditions. The difference between two studies may be due to the impact of pattern of invariance involved in Chen's study; however, this condition was not included in the current study design. As a result, the interactive effect between the proportion of invariance and pattern of invariance could not be detected. Additionally, this finding may also be influence by other factors used in the current design including threshold symmetry, number of indicators, and sample size. Future research may include these factors for further investigation.

Overall, although this study reached an opposite conclusion as compared to Chen's study, both findings implicated that when the proportion of non-invariant items is large, invariance tests may hardly detect a non-invariant instrument. Consequently, invalid group comparison results may be obtained by researchers as the invariance tests fail to detect non-invariance.

Fourth, the number of indicators did not have an appreciable impact on the changes in CFI, RMSEA, and SRMR for lack of both loading non-invariance and threshold non-invariance. This result is in line with previous studies (e.g., Chen, 2007), and it implicates that when applied researchers plan to collect data to conduct measurement invariance testing with multiple-group CFA analysis with ordinal data, it may be not necessary to worry about the size of an instrument contributing to lack of invariance as the number of items did not impact the performance of model fit indices substantially when testing measurement invariance.

Fifth, this study examined two locations of source of non-invariance: item loadings only (metric non-invariance) and item thresholds only (scalar non-invariance).

In tests of loading non-invariance, all three model fit indices were sensitive to 25% loading non-invariance than 50% loading non-invariance. Among the three fit statistics, changes in RMSEA were the biggest for lack of 25% loading invariance, and changes in SRMR were the biggest for lack of 50% loading non-invariance. These findings indicated that RMSEA and SRMR were more sensitive to item loading non-invariance than CFI. The results for RMSEA concurs with Chen's (2007) and Rutkowski and Svetina (2017)'s results, which indicated that ΔRMSEA was able to correctly identify loading non-invariant hypotheses in most conditions. Second, in tests of threshold non-invariance, changes in CFI and RMSEA performed better and were more reliable to identify non-invariance than changes in SRMR. However, considering that unusual negative values of ΔRMSEA were detected for lack of threshold non-invariance, CFI is recommended for use when testing threshold non-invariance (scalar non-invariance).

It is noted that although there were no non-invariant item loadings, rejection rates of all three fit indices for loading non-invariance (metric non-invariance) were high when testing a lack of threshold invariance (scalar invariance) across studied conditions. Results demonstrated that a Type I error (i.e., rejecting invariance falsely) may occur when detecting testing threshold non-invariance. Simply put, applied researchers may make an incorrect conclusion and conclude that the source of the non-invariance is due to loading differences when it is truly from threshold discrepancies.

Sixth, this study examined the relation between levels of threshold symmetry and the performance of fit indices. The finding found that the performance of the fit indices was mixed across levels of invariance and across conditions. Generally, non-invariant symmetric thresholds can be more easily detected by CFI and RMSEA than non-invariant

extreme asymmetric thresholds when threshold non-invariance was examined. This pattern was not observed in SRMR. For lack of loading invariance, the levels of threshold symmetry did not have an appreciable impact on changes in CFI and SRMR. Overall, the performance of RMSEA was more consistent than CFI and SRMR.

5.2 Recommendations for Practice

In conclusion, the results of this study demonstrated that when conducting measurement invariance tests, it is helpful for applied researcher to report changes in CFI, RMSEA, and SRMR instead of only reporting Chi-square difference tests. The findings of this dissertation also found that all three fit indices performed relatively well to detect threshold non-invariance (i.e., scalar non-invariance with ordinal data) across various studied conditions. Among the three fit indices, compared to CFI and RMSEA, SRMR tended to perform sub-optimally under some situations examined here, especially when sample size is small (e.g., 200). For example, as shown in Table 4.10, although there were no non-invariant item thresholds, rejection rates of SRMR for threshold non-invariance (scalar non-invariance) were high using suggested cutoff values for lack of loading invariance (metric invariance) across studied conditions with sample size 200. The high rejection rates of SRMR indicated the high Type I error rates (incorrectly rejecting a true null hypothesis) when applied researchers use SRMR to evaluate measurement invariance.

Regarding changes in RMSEA, negative values were identified in some studied conditions when testing both lack of loading invariance and lack of threshold invariance. For example, the mean difference of RMSEA was -0.002 when testing factor loading non-invariance, given the 8-indicator model with extreme asymmetric thresholds, 50%

non-invariant loadings, and with sample size 1200 (see Table 4.7). The negative results were not common as the more constrained model should perform less well than the less constrained model. Although prior research also found negative ΔRMSEAs in their simulation studies (Rutkowski & Svetina, 2014, 2017), RMSEA is not recommended to use as an evaluation of measurement invariance testing in such cases.

In general, CFI tends to show the best and most stable performance for detecting both lack of loading invariance as well as lack of threshold invariance. Although RMSEA and SRMR have some advantageous properties as fit indices for structural equation modeling, the present studies agree with Chen's (2007) recommendations that using CFI for invariance evaluation first, supplemented by RMSEA and SRMR afterward due to the inconsistent performance of RMSEA and SRMR on several studied conditions.

In addition, previous studies have shown that the magnitude of changes in fit indices is complex as it may be influenced by many factors (Chen, 2007; Mead et al., 2008). The current studies also found that cutoff values in model fit indices need to be used with caution since factors such as sample size per group, proportion of non-invariance, threshold symmetry and source of non-invariance may impact the performance of these model fit indices.

Furthermore, it should be noted that one of interesting findings of the studies is that the rejection rates of all three fit indices for lack of loading invariance were substantially low for models with sample size of 1200, 50% non-invariant loading items, and with both 8 and 16 indicators, as well as both symmetric and extreme asymmetric conditions. These results indicate that it is difficult to detect non-invariance under the above combined conditions when testing metric invariance (loading invariance). Applied

researchers may make a wrong conclusion if their data or models match these studied conditions. More simulation studies are needed to investigate the individual effect or the interactive effect of these studied conditions on performance and rejection rates of CFI, RMSEA and SRMR.

Although it is difficult to propose standards for testing measurement invariance when analyzing ordinal data, it is still useful to provide guidelines derived from these two studies to assist applied researchers for their own research. Table 5.1 reports recommendations regarding the model fit changes assuming metric and scalar invariance. Cutoff values proposed in this study are based on the mean, 5^{th} and 1^{st} average values for CFI or 95^{th} and 99^{th} for RMSEA and SRMR across all study conditions. As CFI and RMSEA are more sensitive to non-invariance in thresholds than loadings, and SRMR is almost equally sensitive to invariance in both loadings and thresholds. Two different cutoff values are recommended for CFI and RMSEA, and one cutoff value is recommended for SRMR.

Specifically, for testing loading invariance (metric invariance) under the conditions studied here, a change of $\leq |\pm 0.003|$ in CFI, supplemented by a change of $\leq |\pm 0.001|$ in RMSEA or a change of ≤ 0.007 in SRMR would indicate metric invariance; for testing threshold invariance (scalar invariance), a change of $\leq |\pm 0.03|$ in CFI, supplemented by a change of $\leq |\pm 0.02|$ in RMSEA or ≤ 0.007 in SRMR would indicate scalar invariance. It is worth noting that all recommended cutoff values reflect the average performance of the three model fit changes across studied conditions. None of the cutoff values perform equally well across all the conditions studied here. In many

conditions present in empirical research situations, it is not known how the criteria perform, and applied researchers should use the indices with caution.

Table 5.1 Recommended model fit cutoff values for different fit measures and invariance levels

Fit Indices	Metric Invariance	Scalar Invariance
CFI	$\leq \lvert\pm 0.003\rvert$	$\leq \lvert\pm 0.03\rvert$
RMSEA	$\leq \lvert\pm 0.001\rvert$	$\leq \lvert\pm 0.02\rvert$
SRMR	≤ 0.007	≤ 0.007

5.3 Limitations and Future Studies

As all simulation studies can only manipulate a limited number of conditions, the current studies have several limitations. First, although conditions in present studies reflected the real-world situation and are simulated based on recommendations from prior research, only a small number of conditions were selected for the two studies. Other potential factors such as unequal sample size per group, model misspecification, model complexity, pattern of invariance, number of groups may be worthwhile to be considered by researchers in their future studies. For example, a study conducted by Rutkowski and Svetina (2017) evaluated the performance of fit indices including Chi-square difference tests, CFI and RMSEA in a large number of groups and varied sample size context using a simulation study. Future studies should continue this line of research, by designing other conditions that may impact the performance of model fit indices.

In addition, the present studies only considered two types of proportion of non-invariance: 25% and 50%. Concerning that the finding of Study 2 is in contrast to Chen's (2007) results, researchers are encouraged to add more proportion levels in their future studies for investigation. For example, a total of five proportion levels of invariance (0%, 25%, 50%, 75% or 100%) were examined in Chen's study (2007), and three proportion

levels of non-invariance (0%, 40% or 60% in 5-item conditions, 0%, 33% or 50% in 6-item conditions) were examined in Rutkowski and Svetina (2017).

A second limitation is that results of second-order chi-square difference test, which is termed as DIFFTEST in Mplus, are not included in this dissertation. While studies have shown that chi-square different test has several limitations (e.g., Chen, 2007; Flora & Curran, 2004; Babyak & Green, 2010), the Chi-square difference test (or DIFFTEST in Mplus with robust estimation) is still widely used by researchers and practitioners. Thus, it is meaningful to understand how non-normal ordered data estimated by WLSMV estimator influences the performance of DIFFTEST. One major difficulty to examine the performance of DIFFTEST is that Mplus Monto Carlo simulation does not support saving DIFFTEST results. As a result, it is time-consuming to save all DIFFTEST results based on various levels of invariance and across studied conditions with many (e.g., 1000) replications when conducting Monte Carlo simulation studies. Although saving DIFFTEST results is tedious, it is still applicable with the assistance of other software packages such as MplusAutomation package in R. In general, it is imperative that future studies may be conducted to guide researchers about the performance of DIFFTEST with ordered categorical data.

Third, model identification was not discussed in the studies. Model specification and identification is a complex issue for multiple-CFA analysis. Invariance testing in the ordered-categorical data is different from in the continuous data as the threshold parameters are involved as a new source of non-invariance, and the factor model is not directly connected to the measured variables anymore (Millsap & Yun-Tein, 2004). However, literature on multiple-CFA analysis with ordered-categorical data is rare

(Millsap & Yun-Tein, 2004). When using M*plus* software program, two parameterizations are offered with WLS estimator family: Delta parameterization and Theta parameterization. The two parameterization methods require constraining invariance differently. Therefore, the performance of fit indices may be influenced by various invariance constraint methods when examining measurement invariance with ordered data. Applied researchers will be beneficial from future studies that aim to examine the relation between model identification and model fit indices with invariance testing using ordered data.

Another limitation of this study is that all results in this dissertation are based on simulated data, empirical data is more likely to reflect the real-world situation and is more complex. Thus, future studies may simulate data based on empirical results or directly use empirical data to validate the study findings.

5.4 Summary and Significance of the Study

In summary, the present studies were an initial step in evaluating the performance of model fit indices (CFI, RMSEA, and SRMR) when measurement invariance is tested in the context of multiple-group CFA analysis with categorical-ordered data. As applied researchers are increasingly aware of the importance of testing measurement invariance, and the prevalence of Likert scales for collecting data, specific recommended guidelines can assist in the evaluation of model fit.

Although some of the findings are in contrast to previous research (e.g., changes in SRMR performance in Chen, 2007), the findings of current dissertation are informative and add to the body of research in the measurement invariance testing literature in a number of ways. It is hoped that the findings of the studies provided here at least may be

a reference for applied researchers and provide useful information to help them conduct their own research. To my knowledge, this is the first study that investigated the performance of fit statistics when data are non-normal and categorical ordered with WLSMV estimator. Additional studies and conditions are needed to examine the performance of fit indices in such settings.

REFERENCES

Alavi, M., Visentin, D. C., Thapa, D. K., Hunt, G. E., Watson, R., & Cleary, M. (2020). Chi-square for model fit in confirmatory factor analysis. *Journal of Advanced Nursing*, 76(9), 2209-2211. https://doi.org/10.1111/jan.14399

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3-23). Hillside, NJ: Lawrence Erlbaum.

Asparouhov, T & Muthén, B. (2006) Robust Chi Square Difference Testing with Mean and Variance Adjusted Test Statistics. MplusWeb Notes: No. 10. http://statmodel.com/download/webnotes/webnote10.pdf

Asparouhov, T., & Muthén, B. (2010). Simple second order chi-square correction. *Mplus technical appendix*, 1-8. https://www.statmodel.com/download/WLSMV_new_chi21.pdf

Asparouhov, T., & Muthén, B. (2018). SRMR in Mplus. *Mplus technical appendix.* https://www.statmodel.com/download/SRMR2.pdf

Babyak, M. A., & Green, S. B. (2010). Confirmatory factor analysis: an introduction for psychosomatic medicine researchers. *Psychosomatic Medicine, 72*(6), 587-597. https://doi.org/10.1097/PSY.0b013e3181de3f8a

Bagozzi, R. P., & Edwards, J. R. (1998). A general approach for representing constructs in organizational research. *Organizational Research Methods*, *1*(1), 45-87. https://doi.org/10.1177/109442819800100104

Bandalos, D. L., & Webb, M. (2005, April). *Efficacy of the WLSMV estimator for coarsely categorized and nonnormally distributed data* [Paper presentation]. American Educational Research Association, Montreal, Canada.

Bandalos, D. L. (2008). Is parceling really necessary? A comparison of results from item parceling and categorical variable methodology. *Structural Equation Modeling: A Multidisciplinary Journal*, *15*(2), 211-240. https://doi.org/10.1080/10705510801922340

Bauer, D. J. (2017). A more general model for testing measurement invariance and differential item functioning. *Psychological Methods*, *22*(3), 507-526. https://doi.org/10.1037/met0000077

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling*, *13*(2), 186-203. https://doi.org/10.1207/s15328007sem1302_2

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238-246. https://doi.org/10.1037/0033-2909.107.2.238

Bergh, D. (2015). Sample size and chi-squared test of fit—a comparison between a random sample approach and a chi-square value adjustment method using Swedish adolescent data. In: Zhang, Q., Yang, H. (eds), *Pacific Rim Objective Measurement* [Symposium] (PROMS) 2014 Conference Proceedings. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-47490-7_15

Browne, M. W., & Cudeck, R. (1992). Alternative Ways of Assessing Model Fit. *Sociological Methods & Research*, 21(2), 230–258. https://doi.org/10.1177/0049124192021002005

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. The Guilford Press.

Bryant, F. B., & Satorra, A. (2012). Principles and practice of scaled difference chi-square testing. *Structural Equation Modeling: A Multidisciplinary Journal*, *19*(3), 372-398. https://doi.org/10.1080/10705511.2012.687671

Bollen, K. A. (1989). *Structural equations with latent variables.* John Wiley & Sons. https://doi.org/10.1002/9781118619179

Boomsma, A. (1985). Nonconvergence, improper solutions, and starting values in LISREL maximum likelihood estimation. *Psychometrika*, 50, 229-242. https://doi.org/10.1007/BF02294248

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: the issue of partial measurement invariance. *Psychological Bulletin*, *105*(3), 456-466. https://doi.org/10.1037/0033-2909.105.3.456

Cabrera-Nguyen, P. (2010). Author guidelines for reporting scale development and validation results in the Journal of the Society for Social Work and Research. *Journal of the Society for Social Work and Research*, *1*(2), 99-103. https://doi.org/10.5243/jsswr.2010.8

Comrey, A. L., & Lee, H. B. (2013). *A first course in factor analysis*. Psychology Press. https://doi.org/10.4324/9781315827506

Cieciuch, J. A. N., & Davidov, E. (2015). Establishing measurement invariance across online and offline samples. A tutorial with the software packages Amos and Mplus. *Studia Psychologica: Theoria et praxis*, *2*(15), 83-99. https://doi.org/10.5167/uzh-170024

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. https://doi.org/10.1080/10705510701301834

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*(2), 233-255. https://doi.org/10.1207/S15328007SEM0902_5

Chou, C. P., Bentler, P. M., & Satorra, A. (1991). Scaled test statistics and robust standard errors for non-normal data in covariance structure analysis: a Monte Carlo study. *British Journal of Mathematical and Statistical Psychology*, *44*(2), 347-357. https://doi.org/10.1111/j.2044-8317.1991.tb00966.x

Chyung, S. Y., Roberts, K., Swanson, I., & Hankinson, A. (2017). Evidence-based survey design: The use of a midpoint on the Likert scale. *Performance Improvement*, *56*(10), 15-23. https://doi.org/10.1002/pfi.21727

Crede, M., & Harms, P. (2019). Questionable research practices when using confirmatory factor analysis. *Journal of Managerial Psychology*. *34*(1), 18-30. https://doi.org/10.1108/JMP-06-2018-0272

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods*, *1*(1), 16-29. https://doi.org/10.1037/1082-989X.1.1.16

Curran, P. J., Bollen, K. A., Paxton, P., Kirby, J., & Chen, F. (2002). The noncentral chi-square distribution in misspecified structural equation models: Finite sample results from a Monte Carlo simulation. *Multivariate Behavioral Research*, *37*(1), 1-36. https://doi.org/10.1207/S15327906MBR3701_01

Ene, M. C. (2020). *Investigating Accuracy of Model Fit Indices in Multilevel Confirmatory Factor Analysis* [Doctoral dissertation], University of South Carolina. https://scholarcommons.sc.edu/etd/6024

DiStefano, C. (2002). The impact of categorization with confirmatory factor analysis. *Structural Equation Modeling*, *9*(3), 327-346. https://doi.org/10.1207/S15328007SEM0903_2

DiStefano, C., & Morgan, G. B. (2014). A comparison of diagonal weighted least squares robust estimation techniques for ordinal data. *Structural Equation Modeling: A*

Multidisciplinary Journal, *21*(3), 425-438. https://doi.org/10.1080/10705511.2014.915373

DiStefano, C. (2016). Examining fit with structural equation models. In K. Schweizer & C. DiStefano (Eds.), *Principles and methods of test construction: Standards and recent advances* (pp. 166–193). Hogrefe Publishing.

DiStefano, C., Liu, J., Jiang, N., & Shi, D. (2018). Examination of the weighted root mean square residual: Evidence for trustworthiness?. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(3), 453-466. https://doi.org/10.1080/10705511.2017.1390394

DiStefano, C., McDaniel, H. L., Zhang, L., Shi, D., & Jiang, Z. (2019). Fitting large factor analysis models with ordinal data. *Educational and Psychological Measurement, 79*(3), 417-436. https://doi.org/10.1177/0013164418818242

Finney, S. J., & DiStefano, C. (2013). Nonnormal and categorical data in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 439–492). IAP Information Age Publishing.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*(4), 466-491. https://doi.org/10.1037/1082-989X.9.4.466

French, B. F., & Finch, W. H. (2006). Confirmatory factor analytic procedures for the determination of measurement invariance. *Structural Equation Modeling*, *13*(3), 378-402. https://doi.org/10.1207/s15328007sem1303_3

French, B. F., & Finch, W. H. (2008). Multigroup confirmatory factor analysis: Locating the invariant referent sets. *Structural Equation Modeling-a Multidisciplinary Journal*, *15*(1), 96-113. https://doi.org/10.1080/10705510701758349

Foldnes, N., & Olsson, U. H. (2016). A simple simulation technique for nonnormal data with prespecified skewness, kurtosis, and covariance matrix. *Multivariate Behavioral Research*, *51*(2-3), 207-219. https://doi.org/10.1080/00273171.2015.1133274

Jackson, D. L., Gillaspy Jr, J. A., & Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: an overview and some recommendations. *Psychological Methods*, *14*(1), 6-23. https://doi.org/10.1037/a0014694

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. https://doi.org/10.1080/10705511.2011.557337

Kim, E. S. (2012). *Testing measurement invariance using MIMIC: Likelihood ratio test and modification indices with a critical value adjustment* [Doctoral dissertation], Texas A & M University. https://core.ac.uk/download/pdf/147230384.pdf

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18*(3), 117-144. https://doi.org/10.1080/03610739208253916

Hoffman, n.d. Measurement Invariance (MI) in CFA and Differential Item Functioning (DIF) in IRT/IFA [PowerPoint slides]. https://www.lesahoffman.com/CLP948/CLP948_Lecture07_Invariance.pdf

Hu, L.-T., & Bentler, P. M. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 76–99). Sage Publications, Inc.

Hu, L. T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*(4), 424-453. https://doi.org/10.1037/1082-989X.3.4.424

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: a multidisciplinary journal*, *6*(1), 1-55. https://doi.org/10.1080/10705519909540118

Li, C. H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior research Methods*, *48*(3), 936-949. https://doi.org/10.3758/s13428-015-0619-7

Lubke, G. H., & Muthén, B. O. (2004). Applying multigroup confirmatory factor models for continuous outcomes to Likert scale data complicates meaningful group comparisons. *Structural Equation Modeling, 11*(4), 514-534. https://doi.org/10.1207/s15328007sem1104_2

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, *34*(2), 183-202. https://doi.org/10.1007/BF02289343

Jöreskog, K. G. (1971). Simultaneous factor analysis in several Populations. *Psychomerika*, *36*(4), 409-426.

Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2015). The performance of RMSEA in models with small degrees of freedom. *Sociological Methods & Research*, *44*(3), 486-507. https://doi.org/10.1177/0049124114543236

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: Guilford.

Lai, M. H., Richardson, G. B., & Mak, H. W. (2019). Quantifying the impact of partial measurement invariance in diagnostic research: An application to addiction research. *Addictive Behaviors*, *94*, 50-56. https://doi.org/10.1016/j.addbeh.2018.11.029

Liang, X., & Yang, Y. (2014). An evaluation of WLSMV and Bayesian methods for confirmatory factor analysis with categorical indicators. *International Journal of Quantitative Research in Education*, *2*(1), 17-38. https://doi.org/10.1504/IJQRE.2014.060972

Liu, Y., Millsap, R. E., West, S. G., Tein, J. Y., Tanaka, R., & Grimm, K. J. (2017). Testing measurement invariance in longitudinal data with ordered-categorical measures. *Psychological Methods*, *22*(3), 486-506. https://doi.org/10.1037/met0000075

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological methods*, *1*(2), 130-149. https://doi.org/10.1037/1082-989X.1.2.130

Maydeu-Olivares, A. (2017). Maximum likelihood estimation of structural equation models for continuous data: Standard errors and goodness of fit. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(3), 383-394. https://doi.org/10.1080/10705511.2016.1269606

Maydeu-Olivares, A. (2017). Assessing the size of model misfit in structural equation models. *Psychometrika*, *82*(3), 533-558. https://doi.org/10.1007/s11336-016-9552-7

Mellenbergh, G. J. (1989). Item bias and item response theory. *International journal of educational research*, *13*(2), 127-143. https://doi.org/10.1016/0883-0355(89)90002-5

Meredith, W. (1964). Notes on factorial invariance. *Psychometrika*, *29*(2), 177-185. https://doi.org/10.1007/BF02289699

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrica*, *58*(4), 525-543. https://doi.org/10.1007/BF02294825

Millsap, E. (2011). Statistical methods for studying measurement invariance. *Taylor & Fransis: Abingdon*, UK.

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate behavioral research*, *39*(3), 479-515. https://doi.org/10.1207/S15327906MBR3903_4

Muthén, B., & Kaplan, D. (1985). A comparison of some methodologies for the factor-analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology, 38,* 171–180. https://doi.org/10.1111/j.2044-8317.1985.tb00832.x

Muthén, B. O. (1993). Goodness of fit with categorical and other non-normal variables, in Bollen, K.A. and Long, J.S. (Eds.): *Testing Structural Equation Models*, pp.205–243, Sage, Newbury Park, CA

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling*, *9*(4), 599-620. https://doi.org/10.1207/S15328007SEM0904_8

Muthén, B. O. (2018). Mplus (Version 8.1) [Computer program]. Los Angeles, CA: Muthén & Muthén.

Muthén, L. K., & Muthen, B. (2010). Mplus 6.0. *Los Angeles, CA: Muthén & Muthén*.

Nevitt, J., & Hancock, G. R. (2004). Evaluating small sample approaches for model test statistics in structural equation modeling. *Multivariate Behavioral Research*, *39*(3), 439-478. https://doi.org/10.1207/S15327906MBR3903_3

Nunnally, J. C. (1967). Psychometric theory. New York, NY: McGraw-Hill

Pavlov, G., Shi, D., & Maydeu-Olivares, A. (2020). Chi-square difference tests for comparing nested models: An evaluation with non-normal data. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(6), 908-917. https://doi.org/10.1080/10705511.2020.1717957

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. https://doi.org/10.1016/j.dr.2016.06.004

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology*, *87*(3), 517-529. https://doi.org/10.1037/0021-9010.87.3.517

Raykov, T., & Marcoulides, G. A. (2012). *A first course in structural equation modeling*. Routledge.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. *Psychological Bulletin*, *114*(3), 552-566. https://doi.org/10.1037/0033-2909.114.3.552

Rensvold, R. B., & Cheung, G. W. (1998). Testing measurement models for factorial invariance: A systematic approach. *Educational and Psychological Measurement, 58*(6), 1017-1034. https://doi.org/10.1177/0013164498058006010

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17*(3), 354-373. https://doi.org/10.1037/a0029315

Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement*, *74*(1), 31-57. https://doi.org/10.1177/0013164413498257

Rutkowski, L., & Svetina, D. (2017). Measurement invariance in international surveys: Categorical indicators and fit measure performance. *Applied Measurement in Education*, *30*(1), 39-51. https://doi.org/10.1080/08957347.2016.1243540

Sass, D. A. (2011). Testing measurement invariance and comparing latent factor means within a confirmatory factor analysis framework. *Journal of Psychoeducational Assessment*, *29*(4), 347-363. https://doi.org/10.1177/0734282911406661

Sass, D. A., Schmitt, T. A., & Marsh, H. W. (2014). Evaluating model fit with ordered categorical data within a measurement invariance framework: A comparison of estimators. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(2), 167-180. https://doi.org/10.1080/10705511.2014.882658

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

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 399–419). Sage Publications, Inc.

Satorra, A., & Bentler, P. M. (2010). Ensuring positiveness of the scaled difference chi-square test statistic. *Psychometrika*, *75*(2), 243-248. https://doi.org/10.1007/s11336-009-9135-y

Satterthwaite, F. E. (1941). Synthesis of variance. *Psychometrika*, *6*(5), 309-316. https://doi.org/10.1007/BF02288586

Savalei, V., Brace, J., & Fouladi, R. T. (2021). We need to change how we compute RMSEA for nested model comparisons in structural equation modeling. https://doi.org/10.31234/osf.io/wprg8

Schmitt, T. A. (2011). Current methodological considerations in exploratory and confirmatory factor analysis. *Journal of Psychoeducational Assessment*, *29*(4), 304-321. https://doi.org/10.1177/0734282911406653

Shi, D. (2016). Resolving three important issues on measurement invariance using bayesian structural equation modeling (BSEM) [Doctoral dissertation], University of Oklahoma.

Shi, D., Lee, T., & Maydeu-Olivares, A. (2019). Understanding the model size effect on SEM fit indices. *Educational and Psychological Measurement*, *79*(2), 310-334. https://doi.org/10.1177/0013164418783530

Shi, D., Song, H., & Lewis, M. D. (2019). The impact of partial factorial invariance on cross-group comparisons. *Assessment*, *26*(7), 1217-1233. https://doi.org/10.1177/1073191117711020

Short, S. D. (2014). *Power of alternative fit indices for multiple group longitudinal tests of measurement invariance* [Doctoral dissertation], University of Kansas.

Sokolov, B. (2019). Sensitivity of goodness of fit indices to lack of measurement invariance with categorical indicators and many groups. *Higher School of Economics Research Paper No. WP BRP*, *86*.

Steiger, J. H., & Lind, J. C. (1980). Statistically based tests for the number of common factors, paper presented at the annual meeting of the Psychometric Society. *Iowa City*, *IA*.

Steiger, J. H. (1998). A note on multiple sample extensions of the RMSEA fit index. *Structural Equation Modeling: A Multidisciplinary Journal*, *5*(4), 411-419. https://doi.org/10.1080/10705519809540115

Tay, L., Meade, A. W., & Cao, M. (2015). An overview and practical guide to IRT measurement equivalence analysis. *Organizational Research Methods*, *18*(1), 3-46. https://doi.org/10.1177/1094428114553062

Thurstone, L. L. (1947). *Multiple-factor analysis; a development and expansion of The Vectors of Mind*. University of Chicago Press.

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), 4-70. https://doi.org/10.1177/109442810031002

VandenBos, G. R. (2014). *APA dictionary of statistics and research methods.* S. Zedeck (Ed.). Washington, DC: American Psychological Association.

Walker, C. M. (2011). What's the DIF? Why differential item functioning analyses are an important part of instrument development and validation. *Journal of Psychoeducational Assessment*, *29*(4), 364-376. https://doi.org/10.1177/0734282911406666

West, S. G., Taylor, A. B., & Wu, W. (2012). Model fit and model selection in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 209–231). The Guilford Press.

Westland, J. C. (2010). Lower bounds on sample size in structural equation modeling. *Electronic commerce research and applications*, *9*(6), 476-487. https://doi.org/10.1016/j.elerap.2010.07.003

Widaman, K. F., & Grimm, K. J. (2014). Advanced psychometrics: Confirmatory factor analysis, item response theory, and the study of measurement invariance. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 534–570). Cambridge University Press.

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281–324). American Psychological Association. https://doi.org/10.1037/10222-009

Woods, C. M., & Grimm, K. J. (2011). Testing for nonuniform differential item functioning with multiple indicator multiple cause models. *Applied Psychological Measurement*, *35*(5), 339-361. https://doi.org/10.1177/0146621611405984

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement*, *73*(6), 913-934. https://doi.org/10.1177/0013164413495237

Xia, Y., & Yang, Y. (2019). RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: The story they tell depends on the estimation methods. *Behavior Research Methods*, *51*(1), 409-428. https://doi.org/10.3758/s13428-018-1055-2

Yu, C. Y., & Muthén, B. O. (2002). Evaluation of model fit indices for latent variable models with categorical and continuous outcomes. [Doctoral dissertation], University of California,

Yuan, K. H., & Bentler, P. M. (2000). 5. Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological methodology*, *30*(1), 165-200. https://doi.org/10.1111/0081-1750.00078

APPENDIX A

SAMPLE MPLUS DATA GENERATION AND DATA ANALYSIS CODE

```
Montecarlo:
   names = u1-u10; !Define the names of the variables.
   generate = u1-u10(4); !Define how many thresholds are needed (k-1)
   categorical = u1-u10;  !Designate variables as ordinal (i.e., ordered categories).
   nrep = 1000; !Specify how many replications per cell.
   seed = 72521; !Seed
   nobs = 150 150; !Specify number of observations.
   ngroups=2;
  results=configural_results_300.dat;
ANALYSIS:
   estimator = wlsmv;
   PARAMETERIZATION=THETA;
MODEL POPULATION:                !Specify the population model
   f1 BY u1-u5@.8;
   f2 BY u6-u10@.8;
   f1@1; f2@1; !Specify factor variances.
   [f1@0]; !Specify factor means
   [f2@0];
   f1 WITH f2@.6; !Specify covariance (correlation) between F1 and F2.
```

```
! thresholds   !These are for extreme asymmetric threshold condition
   [u1$1-u10$1*-1.34
   u1$2-u10$2*-0.84
   u1$3-u10$3*-0.44
   u1$4-u10$4*-0.05];
   u1-u10*0.36  !Specify the uniqueness terms
model population-g2:
MODEL:
f1 BY u1-u5* ; ! Factor loadings all freely estimated, just labeled
f2 BY u6-u10* ;
[u1$1-u10$1*] ;
[u1$2-u10$2*] ;  ! Item thresholds all freely estimated, just labeled
[u1$3-u10$3*] ;
[u1$4-u10$4*] ;
u1-u10@0.36; !
f1@1 f2@1; ! Factor variance fixed to 1 for identification
[f1@0 f2@0]; ! Factor mean fixed to 0 for identification (Mplus forces)
f1 WITH f2* ; ! Factor correlation is freely estimated, just labeled
MODEL g2:
f1 BY u1-u5*;
f2 BY u6-u10*;
[u1$1-u10$1*];
[u1$2-u10$2*];
```

[u1$3-u10$3*];

[u1$4-u10$4*];

u1-u10@0.36;

f1@1;

f2@1;

[f1@0 f2@0];

f1 WITH f2;