An Examination of Measurement Invariance With a Multi-Group Confirmatory Factor Analysis Approach

Ruiqin Gao

Follow this and additional works at: https://scholarcommons.sc.edu/etd

Part of the Educational Psychology Commons

Recommended Citation

This Open Access Dissertation is brought to you by Scholar Commons. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of Scholar Commons. For more information, please contact digres@mailbox.sc.edu.
AN EXAMINATION OF MEASUREMENT INVARIANCE WITH A MULTI-GROUP CONFIRMATORY FACTOR ANALYSIS APPROACH

by

Ruiqin Gao

Bachelor of Arts
Shanxi University, 1998

Master of Arts
Shanxi University, 2002

Master of Education
University of South Carolina, 2017

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

2023

Accepted by:

Christine DiStefano, Major Professor
Jin Liu, Committee Member
Angela Starrett, Committee Member
Dexin Shi, Committee Member

Cheryl L. Addy, Interim Vice Provost and Dean of the Graduate School
DEDICATION

I would love to dedicate my dissertation work to my beloved family, Rujun Ma, Michael Ma, Matthew Ma, and Mia Ma, for their unconditional love and support throughout my graduate life. Thank you for encouraging me when I felt down, thank you for bringing me endless happiness for being just beside me, thank you for being willing to share my happiness and sadness, thank you for giving me enough freedom to pursue my dream, thank you for letting me be myself. Your love matters so much to me and to completing this long academic journey.
ACKNOWLEDGMENTS

I would first like to express my greatest gratitude to my mentor, advisor, and role model, Dr. DiStefano. Thank you for admitting me to this wonderful program and giving me the opportunity to start this amazing academic journey. Thank you for all your unconditional love, understanding, support, guidance, help, trust, and encouragement throughout this journey. Words are not enough to express my gratitude to you. You are the sunshine of my life, directing me to a brighter future.

I would also like to thank Dr. Jin Liu for your support and guidance in my research, study, and life. Thank you for all your trust and encouragement. Thank you for all those research products produced with our hard work together. A lot of beautiful memory with you on this journey. Thankful for meeting you at USC.

I also would like to thank my dissertation committee members, Dr. Angela Starrett and Dr. Dexin Shi. I am very thankful for all your guidance, feedback, insights, advice, and support in completing my dissertation.

I would like to thank all my professors, friends, and colleagues. Thank you for inspiring me and helping me complete this journey in different ways. Thank you for sharing this journey with me.

I would also like to thank my family for all your support, understanding, and trust. It is impossible for me to complete this journey without all the love from you. Thank you with all my heart.
ABSTRACT

This multiple-manuscript dissertation explored the measurement invariance (MI) testing with multiple-group confirmatory factor analysis (MG-CFA) approach from different perspectives. Study 1 explored MI from a theoretical perspective by conducting a systematic review study on MI practices in education. The findings of this study indicated inconsistency in MI practices and showcased the limitations of the MI practices conducted by researchers in the field of education. Study 2 examined MI from an empirical perspective by implementing a cultural MI test of Strengths and Difficulties Questionnaires for elementary school students in the United States and China. The study provided a step-by-step demonstration of how to conduct an MI test appropriately. Study 3 investigated MI from a methodological perspective with a simulation study. This study examined the impact of model size and group size ratio on the sensitivity of fit measures to detect MI. The study found that model size, in combination with group size ratio, affected the power of CFI, RMSEA, and SRMR for identifying the metric or scalar noninvariance. Study 2 and Study 3 originated from the issues with MI practice identified from Study 1 and served as extensions of Study 1. These three manuscripts contributed to the research on MI testing with the MG-CFA approach theoretically, empirically, and methodologically. Overall, this series of studies help researchers gain a better understanding of the application of MI from different perspectives.
# TABLE OF CONTENTS

Dedication........................................................................................................................................ iii

Acknowledgments........................................................................................................................... iv

Abstract............................................................................................................................................... v

Table of Contents................................................................................................................................... vi

List of Tables ......................................................................................................................................... vii

List of Figures ....................................................................................................................................... viii

Chapter 1: Introduction ....................................................................................................................... 1

Chapter 2: Study 1 Examining Measurement Invariance Practices in Education: A Systematic Review .............................................................................................................................. 12

Chapter 3: Study 2 Factor Structure and Cultural Measurement Invariance of the Strengths and Difficulties Questionnaires for Elementary School Students in the United States and China.............................................................. 63

Chapter 4: Study 3 The Impact of Group Size Ratio and Model Size on the Sensitivity of Fit Measures in Measurement Invariance Testing: A Monte Carlo Simulation Study ......................................................................................... 98

Chapter 5: Conclusion .......................................................................................................................... 147

Bibliography ......................................................................................................................................... 155
LIST OF TABLES

Table 2.1 Checklist for MG-CFA review .................................................................54
Table 2.2 Data screening results from MG-CFA review articles .............................55
Table 2.3 Estimation and fit indices used in MG-CFA review articles ......................56
Table 2.4 Model fit indices used in MG-CFA review articles ................................58
Table 2.5 Fit statistics criteria used in MG-CFA review articles ...............................59
Table 2.6 Percentage of MI levels established in MG-CFA review articles ................59
Table 2.7 Pearson correlations of model fit criteria and sample characteristics with levels of invariance ........................................60
Table 3.1 Demographic characteristics of participants ...........................................93
Table 3.2 CFA model fit statistics ..........................................................................94
Table 3.3 Three-factor CFA results: standardized factor loadings for group samples .................................................................95
Table 3.4 Correlations among teacher-reported SDQ factors .................................96
Table 3.5 MG-CFA mode fit across the U.S. and Chinese samples .........................96
Table 4.1 Fixed parameters in the population model .............................................129
Table 4.2 Simulation design factors and manipulations ........................................129
Table 4.3 Power rates of $\Delta$CFI, $\Delta$RMSEA, and $\Delta$SRMR for detecting metric invariance ..........................................................130
Table 4.4 Power rates of $\Delta$CFI, $\Delta$RMSEA, and $\Delta$SRMR for detecting scalar invariance ..........................................................131
LIST OF FIGURES

Figure 2.1 Number of studies on MI from 1991-2021.............................................................61

Figure 2.2 Flow chart of the study selection process...............................................................62

Figure 3.1 The three-factor structure of teacher-reported SDQ.............................................97

Figure 4.1 The power rates of ΔCFI for detecting metric noninvariance........................................132

Figure 4.2 Power rates of ΔRMSEA for detecting metric noninvariance........................................135

Figure 4.3 Power rates of ΔSRMR for detecting metric noninvariance........................................138

Figure 4.4 Power rates of ΔCFI for detecting scalar noninvariance..............................................141

Figure 4.5 Power rates of ΔRMSEA for detecting scalar noninvariance........................................144
CHAPTER 1

INTRODUCTION

While measuring instruments are widely used in social science, researchers need to provide strong support for the validity of the instruments. Validity refers to the meaningfulness of research components. In other words, an instrument has validity if it accurately measures the construct it is intended to measure (Drost, 2011). Among different types of validity, construct validity is most commonly discussed to provide support for an instrument. Construct validity refers to the degree of agreement between a theoretical concept and a specific measuring instrument or procedure (Fitzner, 2007). Measurement invariance is commonly used to provide evidence of construct validity. This evidence is especially important when research involves samples from diverse groups (e.g., language, gender, and ethnic backgrounds) (González et al., 2017). If researchers mistakenly assume that members from different groups ascribe the same meaning to the items on the measures, the conclusion about the difference in mean scores in the construct across groups is likely to be biased.

Measurement invariance (MI) refers to the psychometric equivalence of the construct of interest across groups or measurement occasions (Cheung & Rensvold, 2002; Putnick & Bornstein, 2016). MI examines whether scores from the operationalization of a construct have the same meaning across conditions, such as time points, administration method, and population subgroups (Meade & Lautenschlager, 2004) and is an essential property of an instrument (Schmitt & Kuljanin, 2008; Steenkamp & Baumgartner, 1998;
Vandenberg & Lance, 2000) as it is a pre-condition for comparing the constructs across groups or conditions (Borsboom, 2006; Davidov, 2011; Putnick & Bornstein, 2016). The lack of MI may imply that the same instrument measures different constructs across groups, or the same construct may have different meanings for different groups (Cieciuch et al., 2019). In other words, assessing the actual differences in the latent construct arises from the differences in the observed scores rather than group or time differences.

A common approach for investigating MI is with confirmatory factor analysis (CFA) frameworks (Kim & Yoon, 2011; Liu et al., 2017). CFA is a statistical framework used to examine the association between a set of observed variables and unobservable constructs (Brown, 2015; Claxton et al., 2015). Under this framework, multiple-group confirmatory factor analysis (MG-CFA) is used to evaluate whether the measured indicators reflect the same construct on the same scale in different groups. With the MG-CFA method, researchers can assess the MI of an instrument by fitting CFA models separately to the data obtained from each group by imposing across-group equality constraints on specific model parameters (Bauer, 2017). MI testing with MG-CFA involves a series of sequential tests, including configural invariance, metric invariance, scalar invariance, and residual invariance tests, where successive tests impose restrictions or equality restraints on model parameters. Restraints that result in an acceptable fit are considered invariant and guide what comparisons should be made. Configural invariance signifies that different groups used equal conceptual groundings to interpret the items (Cheung & Rensvold, 2002; Schmitt & Kuljanin, 2008; Vandenberg & Lance, 2000). Metric invariance indicates that the strength of the relationship between items and an underlying construct is identical across groups (Cheung & Rensvold, 2002; Schmitt &
Kuljanin, 2008; Vandenberg & Lance, 2000). The scalar invariance demonstrates that the origin of the item score on the latent variable is the same across groups (Meredith, 1993; Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000). The residual invariance indicates that the level of measurement error in the items (i.e., variance unexplained by the factors) is the same across groups (Jung & Yoon, 2016).

Different model fit indices can be used for evaluating the MI levels. Here, levels were examined by comparing two nested models (i.e., more restricted model with fewer free parameters and less restricted model with more free parameters). Chi-square tests have been used to determine the difference between baseline and more restricted models (Chen, 2007; Putnick & Bornstein, 2016); however, as the chi-square likelihood ratio is sensitive to the departures from multivariate normality and is always large and statistically significant with complex models and/or large samples (Chen, 2007). Alternative model fit indices, including Comparative Fit Index (CFI), Root Mean Square Error Approximation (RMSEA), and Standardized Root Mean Square Residual (SRMR) were recommended as a supplement for evaluating MI. The configural invariance was evaluated with the recommended criteria of these fit measures. The change in $\chi^2$, CFI, RMSEA, and SRMR between nested models (e.g., configural vs. metric, metric vs. scalar) are suggested for testing the establishment of the rest of the MI tests.

Given the importance of MI testing and the development of statistical techniques for MI testing, the number of studies on MI has been increasing. A search of the Education Resources Information Center Database (ERIC) with delimiters “Measurement invariance” showed that the number has increased since 1991 (n=3) up to 2021(n=4,506). While MI is broadly applied in research, it is necessary to explore MI from different
perspectives. First, it is necessary to review studies on MI to document how researchers apply MI tests in practice and whether there are issues with MI practicing and reporting empirically or methodologically. Second, if issues are identified, it is necessary to further explore the relevant issues by conducting empirical research with data collected from real-life situations or simulation studies with simulated data.

This dissertation consists of three studies examining the measurement invariance (MI) practice with the MG-CFA approach. The first study involved a systematic review of the existing application of MI with the MG-CFA approach in education. The second study illustrates the practice procedures in a non-technical way of how to test MI with the MG-CFA approach with an illustrative example. The study examined the cultural MI in the Strengths and Difficulties Questionnaires (SDQ) for elementary school students in the United States and China. The third study is a simulation study of the impact of group size ratio in combination with model size on the sensitivity of the fit measures in MI testing. The purpose of this collective work is to examine the MI with an MG-CFA approach from different perspectives, including theoretical, empirical, and methodological perspectives.

The first study in this dissertation is a systematic review of the MI with MG-CFA approach in education. As sub-groups of the population in education are, in most cases, heterogeneous, comparisons are often made to identify the similarities and differences across groups of students in education (e.g., gender, race, and grade level), it is necessary to examine the MI of the instruments prior to the cross-group comparison to make sure that instruments measure the constructs with the same meaning across groups or conditions. If MI is established, the differences in the observed scores may accurately
reflect the true differences in latent constructs being measured. With the number of studies on MI dramatically increasing, it is essential to summarize how MI has been applied in past research in education. Two previous studies were conducted on MI with the traditional review method (Schmitt & Kuljanin, 2008; Vanderberg & Lance, 2000). However, the traditional review has been criticized for lack of reliability and validity evidence and research bias (Grant & Booth, 2009). Therefore, this study conducted a systematic review of the MI studies in education to summarize the reporting practices of MI with the MG-CFA approach, identify areas for further improvement in the MI application, build MI reporting guidelines, and provide directions for future MI practices in education. Overall, this study explored MI from a theoretical perspective. This review study lays the theoretical foundation for further exploration of MI in study 2 and study 3.

The review results of Study 1 identified issues with MI practice and developed guidelines for practicing and reporting appropriately. Specifically, the review results showed that only 28.8% of the MI tests were conducted with categorical data (i.e., data with items of 2 to 4 levels), indicating that applying MI tests with categorical data is uncommon. Therefore, there was a need to illustrate to researchers how to implement a MI test with categorical data by providing an empirical example in education. Study 2 examined the cultural measurement invariance in the teacher-reported Strengths and Difficulties Questionnaires across the United States and China. The Strengths and Difficulties Questionnaire (SDQ) is a brief screening instrument designed to assess the behavioral attributes of children aged from 4 to 17 (Goodman, 2001). It is one of the most commonly used screening instruments worldwide (Gao et al., 2013; Guo et al., 2018; Mellor et al., 2016; Obel et al., 2004). As the SDQ is widely used, researchers
investigated various psychometric properties of the SDQ, such as MI. The previous research investigated the MI of SDQ across different groups, including informants, gender, age, race/ethnicity, income, level of education, parents, language versions, settings, and time (Chiorri et al., 2016; DeVries et al., 2017; Gomez & Stavropoulos, 2020; He et al., 2013; Janitza et al., 2020; Liang et al., 2019; Murray et al., 2021; Rogge et al., 2018; Van de Looij-Jansen et al., 2011). No previous study has examined the cultural MI of the teacher version of the SDQ across China and the U.S. within elementary school contexts. Therefore, this study aimed to explore the cultural MI of the teacher version of the SDQ with samples of elementary school children from both China and the U.S.

The illustration provides researchers with a step-by-step procedure for applying MI testing with categorical data. This study incorporated the guidelines of MI testing developed in Study 1. Specifically, the theoretical framework provided by Study 1 and the appropriate MI practices identified from Study 1 informed the methods employed in this study. In turn, Study 2 served as an empirical supplement to Study 1.

In addition, Study 1 found that different model sizes and group size ratios were involved in the research on MI testing. However, the suggested cutoff criteria of the fit measures were commonly used by researchers regardless of group size ratios and model sizes. The findings initiated the necessity of exploring whether group size ratios in combination with model size may affect the sensitivity of fit measures used for assessing MI. Previous research investigated the changes in model fit statistics for detecting MI when different model sizes (Cheung & Rensvold, 2002; French & Finch, 2006; Meade et al., 2008) or group size ratios (Chen, 2007; Yoon & Lai, 2018) were involved. These
previous studies examined the impact of model sizes or group size ratios on the performance of fit measures in detecting MI. The effect of the group size ratio in combination with the model sizes on the power for detecting MI remained to be studied. The study’s findings provided implications for applied researchers regarding the appropriateness of using the commonly used fit measures (i.e., $\Delta$CFI, $\Delta$RMSEA, $\Delta$SRMR) to evaluate MI with varying model sizes and group size ratios. This study served as an extension of Study 1, from which research questions arose. Besides, this study explored MI from a methodological perspective, serving as a supplement to both Study 1 and Study 2.

These three manuscripts contributed to the research on the MG-CFA in education as each manuscript explored a unique aspect of MI, including a review of the research on MI practices in education, an illustration of how to apply MI in empirical research with an example, and a simulation study on the impact of group size ratio in combination with model size on the sensitivity of fit measures used for identifying MI. The systematic review of MI in Study 1 informed researchers of the common MI practice issues and areas to address. The study served as a starting point for exploring MI from a theoretical perspective. The empirical analysis in Study 2 used data collected in real educational contexts to showcase to researchers the detailed procedures for implementing MI, which were identified through the systematic review in Study 1. This study explored MI from an empirical perspective. Study 3 examined the methodological issue identified from Study 1. This study examined the effect of group size ratio in combination with the model size on the sensitivity of the fit measures in MI. The findings of this study provided methodological implications for researchers using alternative fit indices to evaluate MI
with the MG-CFA approach. Overall, this series of studies help researchers gain a better understanding of the application of MI from different perspectives.

References


CHAPTER 2: STUDY 1

EXAMINING MEASUREMENT INVARIANCE PRACTICES IN EDUCATION:
A SYSTEMATIC REVIEW

Abstract

There has been a great increase in the number of studies using measurement invariance (MI) in education. Despite rapidly growing application, the practices of MI in education have yet to be systematically examined. This paper conducted a systematic review of MI studies using PRISMA guidelines. A total of 91 studies were collected from the Education Resources Information Center Database (ERIC) between January 2022 to June 2022. A total of 63 studies were included for review, and 129 MI tests were included in these 63 studies. Three major elements (i.e., data screening, estimation methods and fit indices, and measurement invariance testing) were reviewed to examine the practice of MI testing. Results indicated inconsistency in MI practices and reporting in the previous studies and showcased the limitations of the MI practices conducted by researchers in the field of education. The study provided implications regarding what is expected to report in MI-related studies and what is needed for future research reports.

Keywords: measurement invariance, systematic review, education
Introduction

Comparisons are often made to identify similarities and differences across groups in education. Comparative research usually involves using measurement instruments to assess different aspects of subjects’ behaviors. However, before groups can be compared, researchers need to consider whether instruments measure the constructs with the same meaning across conditions or groups, especially when instruments were developed to assess constructs that are not directly observable, such as attitudes, beliefs, intentions, motives, etc. (Ployhart & Oswald, 2004). If researchers do not examine the measurement invariance of the instruments prior to conducting comparisons, the potential group differences and subsequent associations between the construct and other variables may be due to measurement inaccuracy (Adolf et al., 2014).

Examining responses to determine if groups interpret constructs in the same way is referred to as measurement invariance (MI). MI examines whether scores from the operationalization of a construct have the same meaning across conditions, such as differences in time points, administration method, and population subgroups (Meade & Lautenschlager, 2004). With a questionnaire, MI is tested by evaluating whether the items on a multi-item scale relate to the construct(s) in the same way for all individuals across conditions (e.g., time or groups) (Bauer, 2017). The presence of MI is a prerequisite for comparing psychological constructs across conditions (Borsboom, 2006; Davidov, 2011; Meredith & Teresi, 2006). The lack of MI implies that the same instrument does not measure the same underlying constructs across conditions (Cieciuch et al., 2019). Comparison of groups without establishing MI can lead to problems such as underestimation or overestimation of group differences in item-score means (Jones &
Gallo, 2002), sum-score means (Jeong & Lee, 2019), and regression parameters in structural equation models (Guenole & Brown, 2014).

With the development of statistical techniques for MI testing, MI has gained increasing attention in different social sciences. The literature on testing MI has rapidly increased since 1990 (Bauer, 2017) as more MI tests were applied within a structural equation modeling framework (e.g., Cheung & Rensvold, 1999; Rensvold & Cheung, 1998; Steenkamp & Baumgartner, 1998). The Education Resources Information Center Database (ERIC) with delimiters “Measurement invariance” was examined over a 20-year span from 1991 to 2021. The investigation showed a dramatic increase in the number of studies on MI from 1991 to 2021. As seen in Figure 2.1, only three studies on MI were published in 1991. This number dramatically increased to 1,541 in 2015 and reached 4,506 in 2021. While MI has been broadly applied, the reporting conventions on MI are still in flux and inconsistent (Putnick & Bornstein, 2016). The lack of consistent MI testing and reporting guidelines may lead researchers to gain inaccurate information about the constructs and the groups they study. Therefore, it is essential to review the current practices for testing and reporting MI and discuss the guidelines for appropriate practices.

Previous studies reviewed the use of MI to evaluate the practice of MI in different disciplines. For example, Vandenberg and Lance (2000) reviewed the research on MI in a CFA framework in organizational research. This study reviewed fourteen papers published between 1971 to 1998 to examine how researchers tested various aspects of MI, including which level of tests were evaluated, the sequence in which multiple tests were evaluated, the selection of the model fit indices for the overall model fit, the test of
the differences between models, and reference indicator selection. However, in the review, inconsistencies existed in which invariance tests should be undertaken for a thorough examination of MI, the particular sequence of the tests to be conducted, and the frequency with which each test was discussed. Finally, Vanderberg and Lance (2000) provided guidance to illustrate the steps which should be taken when testing invariance.

Schmitt and Kuljanin (2008) reviewed 75 empirical MI studies published between 2000 and 2007 using the MG-CFA approach. This paper reviewed the types of invariance considered, the content area addressed by the measures involved, the number of groups compared, and whether the measures have been translated into a different language from the original one. This review paper added to the literature by providing more information about the actual practices of researchers who address the issue of MI. For example, more studies tested scalar invariance, partial invariance, and the differences in factor means than what was found in the literature reviewed by Vanderberg and Lance (2000). Few studies tested the significance of the difference in the variance-covariance matrices.

Despite the contribution made to the existing literature on MI, these previous studies have their limitations. First, as all the studies reviewed were conducted before 2007, researchers’ practices should have changed with improvements in computing and methodological developments. Notably, there has been a dramatic increase in the application of MI, suggesting that it is necessary to provide updated research on MI. Therefore, the current study explores the application of MI in the most updated research conducted from January 2022 to June 2022. Second, the study by Vandenberg and Lance (2000) focused on organizational research, while Schmitt and Kuljanin (2008) explored MI without focusing on any specific discipline. As each discipline has its own common
cognitive or social rationale that defines its boundaries (Becher & Trowler, 2001), the way that the research is approached may differ by area.

This review will focus on MI in the education literature base. As sub-groups within populations in the research field of education are often heterogeneous, these groups differ from one another regarding the measurement or structure parameters. Hence, when instruments were developed to measure constructs of interest in these areas, it was critical to assess the MI of the instruments to ensure that the instruments functioned equally across groups. As no previous studies were conducted on the MI of instruments in educational areas, the current study is to fill the gap by focusing on the application of MI in education. Third, the previous review studies were conducted with traditional methods of review (Schmitt & Kuljanin, 2008; Vanderberg & Lance, 2000). However, traditional literature reviews have been critiqued for their lack of scientific rigor, a lack of reliability and validity evidence, and the extent of research bias in producing evidence-based knowledge (Grant & Booth, 2009).

A systematic review approach has been promoted as a method to improve the quality and transparency of literature reviews by reducing biases and omissions (Tranfield et al., 2003). Relative to traditional literature reviews, a systematic review has the advantage of adopting a replicable, scientific, and transparent process and providing unbiased and objective searches (Liberati et al., 2009; Tranfield et al., 2003). The systematic review, which involves the synthesis of findings, adopts “a replicable, scientific and transparent process, in other words, a detailed technology that aims to minimize bias through exhaustive literature searches by providing an audit trail of the reviewers’ decisions, procedures, and conclusion.” (Tranfield et al., 2003, p.209). To
date, no review study of MI in education has been conducted. This may be especially interesting in education, as comparing cognitive outcomes across groups is often a major goal of large-scale educational assessment studies (Robitzsch & Lüdtke, 2022). Given the situation, the current study presented a systematic review of data from studies that have tested MI in education. More specifically, a systematic review was conducted by following the protocol of Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) to increase the transparency and accuracy of literature reviews (Liberati et al., 2009). A comprehensive systematic review of the application of MI in the educational areas against the PRISMA checklist would contribute to having a better understanding of the practice of MI.

Theoretical Framework

Measurement Invariance Testing

MI has been examined within the latent variable modeling framework. Various approaches have been developed to test MI. Multiple-group confirmatory factor analysis (MG-CFA) fits CFA models separately to the data obtained from each group by imposing across-group equality constraints on specific model parameters (Bauer, 2017). The multiple-group item response theory (MG-IRT) examines the possibility of selecting a specific item level, given a score on the latent construct and a specific group membership (D’Urso et al., 2022). The multiple-indicator multiple-cause (MIMIC) model is a special case of SEM and is composed of two parts: a measurement model which defines the relationship between indicators and a latent variable (established at the CFA stage) as well as a structural model that specifies the direct effects of the covariates on one or more item responses and latent factors (Jöreskog & Sörbom, 1996). While all these approaches
allow for detecting differential item functioning within multi-item scales, they have advantages and limitations. For example, MG-CFA and MG-IRT examine the invariance of all model parameters across levels of one categorical grouping variable. The MIMIC model allows for examining the invariance of only a subset of model parameters across levels of categorical and continuous individual difference variables (Bauer, 2017). Among these methods, MG-CFA is the most common method for measurement invariance testing (Vandenberg & Lance, 2000) and will therefore be the focus of this study.

The MG-CFA approach explores the underlying latent structure among a set of observed variables across different groups or occasions within the CFA framework. CFA is a measurement model that tests the association between observed variables (e.g., items, indicators, measures, manifest variables) and unobservable concepts (e.g., factors and latent variables) (Claxton et al., 2015; Floyd & Widaman, 1995). Under an MG-CFA approach, MI is assessed by fitting CFA models separately to the data obtained from each group by imposing across-group equality constraints on specific model parameters (Bauer, 2017). If the model fit is adequate, we can identify the similarities and differences between factor structures and parameter estimates by imposing equality constraints on model parameters.

Researchers are interested in assessing whether 1) the same measurement model is held across groups, including structure patterns (i.e., the relationship between items in a measure), factor loadings (i.e., the strength of the relationship between items and an underlying factor), intercepts/thresholds (i.e., each item’s original value), residual variances (i.e., variance unexplained by the latent factors), and 2) the same structural
model is used across groups, including factor variance and covariance, factor means, or regression coefficients among factors (Vandenberg & Lance, 2000). This study will focus on the measurement model, as structural invariance cannot be assessed unless MI is established (Cyders, 2013).

With MG-CFA, a series of increasingly restrictive models are fitted to test different levels of MI. Typically, an MI investigation comprises four hierarchical levels of invariance, including configural invariance, metric invariance, scalar invariance, and residual variance invariance. Configural invariance is the least restrictive level. In this model, the number of factors and the correspondence between factors and indicators are identical across different conditions, with all parameters being freely estimated in each group (Cheung & Rensvold, 2002; Schmitt & Kuljanin, 2008). Configural invariance indicates that participants from different groups interpret the items based on the same theoretical framework. Metric invariance assumes configural invariance with the additional meaning that the strength of the association between items and latent factors is equivalent across different conditions. Unstandardized factor loadings (i.e., the coefficient of each indicator) were constrained to be the same across conditions in the metric invariance model (Cheung & Rensvold, 2002; Schmitt & Kuljanin, 2008). The metric invariance indicates that the constructs measured by the instruments are manifested in the same way in each group (Kline, 2015). The scalar invariance assumes metric invariance. Scalar invariance indicates that participants from different conditions with the same level on the factor obtain the same score on the indicator (Kline, 2015). The scalar invariance requires the unstandardized intercept to be equal across different conditions (Meredith, 1993; Steenkamp & Baumgartner, 1998; Vandenberg & Lance,
The residual variance is the highest level of invariance and assumes scalar invariance. It examines whether the item variance unexplained by the factors differs across groups (Jung & Yoon, 2016). It also requires equal residual variance and covariance across different conditions. The establishment of the residual invariance implies that the indicators measure the same factor in different conditions with same degree of precision (Kline, 2015).

If full invariance is not established, partial invariance is established by releasing constraints of some item parameters. Among these four MI, the configural and metric invariance are pre-conditions for the equality of the structural relationship between the underlying construct across different conditions. The scalar invariance is a necessary condition for latent mean comparison across conditions (Ployhart & Oswald, 2004). A minimum of two items’ factor loadings and intercepts should be equal across groups for latent means comparison (Lomazzi & Seddig, 2020).

**Model Fit Indices**

Assessing the model fit for each group separately to examine the factorial validity is suggested prior to conducting MI to ensure the factorial validity of the instrument (Sass et al., 2014). After obtaining an adequate model for each group separately, a test of configural invariance is conducted to obtain a baseline model for model comparison purposes. The models for individual samples and the configural invariance models were tested with different model fit indices along with the traditionally recommended cut-off values (e.g., Hu & Bentler, 1999). The commonly used fit indices include the overall Chi-square test of exact fit, Root Mean Square Error of Approximation (RMSEA), Comparative Fit Index (CFI), Tucker-Lewis Index (TLI, or the Non-Normed Fit Index), and
Standardized Root Mean Square Residual (SRMAR) (DiStefano & Hess, 2005; Jackson et al., 2009; Kline, 2015).

In cases of large sample sizes and or a complex model, chi-square is too sensitive to retain an acceptable model, and additional fit indices are often used in conjunction with a significant value (Brown, 2015). RMSEA was used to estimate the lack of fit between the population data and the model estimates. RMSEA values at or below .05 indicate a close model fit, and values between .05 and .08 indicate an adequate fit (Marsh et al., 2004). CFI compares a proposed model’s fit to an independent model, with a value greater than 0.95 indicating a good model fit (Kline, 2015) and a value greater than 0.90 showing an acceptable model fit (Hu & Bentler, 1999). TLI is used to test the improvement of model fit per degree of freedom of the target model over the independent model. A TLI value greater than 0.90 or 0.95 denotes a close or good model fit (Hu & Bentler, 1999; Little, 2013). SRMR estimates the amount of error remaining in the variance-covariance matrix, with values under .08, indicating an acceptable fit (Hu & Bentler, 1999). These model fit indices are typically used to assess whether the proposed model fits a set of data; however, in the context of MI, these indices are used to evaluate the configural invariance as it serves as the baseline model for MI testing.

Two models are nested if the parameters of the more restricted model include some of the parameters of the less restricted model (Bentler & Bonett, 1980). For testing the establishment of the metric, scalar, and residual invariance, the model fit of the two nested models (i.e., configural vs. metric invariance, metric vs. scalar invariance, and scalar vs. residual invariance) is compared. When two models are nested, the more restricted model includes fewer free parameters to estimate than found in the less
restricted model (Savalei et al., 2023). A variety of goodness-of-fit indices were proposed. The chi-square difference test ($\Delta \chi^2$) compared the constrained to the unconstrained nested models by specifying the fixed and free parameters (Byrne et al., 1999; Reise et al., 1993). However, the chi-square difference test in the large sample size may be statistically significant (Chen, 2007; Cheung & Rensvold, 2002). Alternative fit indices (i.e., $\Delta$ CFI, $\Delta$ RMSEA and $\Delta$ SRMR) were also used to evaluate MI. As approximate fit indexes were less affected by group size and the number of factors than the chi-square difference test in large samples, the use of $\Delta$CFI should be preferred over the significance test in large samples (Chen, 2007; Cheung & Rensvold, 2002; Mead et al. 2008). Besides, as $\Delta$CFI is not correlated with overall goodness of fit indices and is independent of the model complexity and sample size, it is recommended for testing MI in CFA models (Cheung & Rensvold, 2002). $\Delta$CFI less than or equal to .01 are suggested (Cheung & Rensvold, 2002). Meade et al. (2008) suggested $\Delta$CFI values less than or equal to .002 when group sizes are very large. As RMSEA was not affected by the number of items per factor or the number of factors in the model, it is often included with MI tests (Cheung & Rensvold, 2002). SRMR was also recommended for MI tests as it is more sensitive to non-invariance in loadings than intercepts or residual variances (Chen, 2007). When sample sizes are small ((Total N≤300), group sizes unequal across comparison groups and the pattern of noninvariance is uniform, criteria of $\Delta$CFI ≤.005 and $\Delta$RMSEA ≤.010, and $\Delta$SRMR ≤0.025 were suggested for testing loading invariance. The criteria of $\Delta$CFI ≤.005, $\Delta$RMSEA ≤.010, and $\Delta$SRMR ≤0.005 were suggested for testing intercept or residual invariance (Chen, 2007). When the sample size is adequate (Total N>300), group size is equal across groups, and the lack of invariance is mixed, more
stringent criteria were suggested. Under more stringent conditions, $\Delta CF_{I} \leq 0.010$, $\Delta RM_{SEA} \leq 0.015$, and $\Delta SR_{MR} \leq 0.30$ were suggested as criteria for evaluating metric invariance and $\Delta CF_{I} \leq 0.010$, $\Delta RM_{SEA} \leq 0.015$, and $\Delta SR_{MR} \leq 0.010$ for scalar and residual invariance (Chen, 2007). $\Delta CF_{I} \leq -0.02$ and $\Delta RM_{SEA} \leq 0.03$ were suggested for metric invariance, and $\Delta CF_{I} \leq -0.01$ and $\Delta RM_{SEA} \leq 0.01$ for scalar invariance in cases where number of groups are large (Rutkowski & Svetina, 2014). In sum, approximate fit indexes are more appropriate than chi-square significance tests to establish MI when sample sizes are large (Chen, 2007; Cheung & Rensvold, 2002; Meade et al., 2008). However, the approximate fit indexes may not perform well with misspecified models with ordinal data (Sass et al., 2014). The application of the rules depends on the research context.

In addition to the above-mentioned comparative and absolute indices, information-theoretic indices, including Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC), are also used to compare competing models and make a trade-off between model fit and model complexity. A lower AIC or BIC value indicates a better model (Van de Schoot et al., 2012). Using these information criteria in MI is relatively less common than the fit measures such as CFI and RMSEA but could provide unique information, especially if continuous data are analyzed (Cao & Liang, 2022).

**Factors Affecting Measurement Invariance**

As sample size affects the power of the study, it also affects the establishment of MI using absolute model fit such as $\chi^2$. When the sample size increases, the chi-square differences could be statistically significant even if the absolute differences in parameter
estimates are trivial (Cheung & Rensvold, 2002). Thus, testing MI with a large sample with $\chi^2$ as the only criterion may lead to the over-rejection of the MI if the imposition of cross-group equality constraints makes relatively little difference in model fit (Kline, 2015). Although changes in alternative fit indices (AFIs) are less sensitive to sample size (Cheung & Rensvold, 2002), measures of absolute model fit such as RMSEA may over-reject true invariance models when the sample size is less than 100 (Chen et al., 2008). Therefore, the sample size plus the choice of model fit criteria may affect the researchers’ decision on the MI. Thus, the current review studies examined the association between the sample size and the level of MI established and the relationship between the choice of model fit criteria and the level of MI identified.

The power of significance tests in MC-CFA is also affected by total sample sizes. The simulation studies by Meade and Bauer (2007) found that the power to detect the estimated factor loading differences was high for total sample sizes of 400 but low for total sample sizes of 100. The power differed significantly by the psychometric properties of the data (e.g., the degree of factor overdetermination, the level of indicator commonalities) for sample sizes of 200. Based on these previous findings, we examined whether the sample size is associated with achieving a certain level of MI.

The previous studies also explored whether the number of groups is associated with identifying MI. The simulation study by Rutkowski and Svetina (2014) suggested that the $\Delta$CFI decreased and $\Delta$RMSEA increased as the number of groups increased from 10 to 20 for metric MI testing. Therefore, they recommended a slightly more liberal criterion of $\Delta$CFI and $\Delta$RMSEA for metric MI testing. Based on these findings, the
current study examined whether the number of groups is associated with establishing a specific level of MI.

In educational research, data often have a hierarchical multi-level structure (Jak et al., 2013), such as data from children nested in the classrooms and schools, members nested in organizations, teachers nested in the schools, or parents nested in the community. The data collected from the same cluster should be considered dependent, as participants from the same cluster might share similar characteristics. Ignoring nesting may increase model misfit, bias parameter estimate, and reduce the standard errors of parameter testing (Hox, 2010; O’Connell & McCoach, 2008; Pornprasertmanit et al., 2014). The level of dependence was interpreted by the intraclass correlation coefficient (ICC), which indicates how strongly cases within the same cluster are interdependent (Kline, 2015). When ICC >.50, the chi-square statistic is inflated, unstandardized parameter estimates are overestimated, and standard errors are underestimated (Julian, 2001; Pornprasertmanit, et al., 2014). Ignoring the multilevel nature of nested data leads to underestimated standard error at the within-level, inflated type I errors, and inaccurate conclusions with biased parameter estimates due to the violation of independence (Snijders & Bosker, 2012). Within the context of MI, the invariant model is more likely to be rejected and concluded to be noninvariant when multi-level observations are not considered (Kim et al., 2012). Thus, the current study examined whether the MI tests took dependencies in the data due to the nesting structure into account to correct standard errors and parameter tests.

The present study will focus on four main objectives: (a) to summarize the reporting practices of MI within the CFA framework; (b) to identify areas that need
improvement when implementing MI; (c) to build MI reporting guidelines; (d) to provide directions for future MI practice in educational areas.

**Methods**

A systematic review of MI studies in education was performed to identify practices used in conducting and reporting MI with the MG-CFA approach. The reporting checklist of PRISMA (Liberati et al., 2009) was followed. All the papers on MI published from January 2022 to June 2022 in educational journals were systematically searched. A protocol was developed to document the analysis method and inclusion criteria. Education Resources Information Center Database (ERIC) was utilized to search for research on MI containing the term “Measurement invariance/ MI” or “Multiple-group Confirmatory Factor Analysis/ MG-CFA” in their titles, abstracts, and/or keywords. The search was limited to year (Year 2022), month (1-6), language (English), research categories (psychology and education), and the type of publication (peer-reviewed empirical papers).

The title, abstract, keywords, author’s name and affiliations, journal name, and publication date were exported to an Excel spreadsheet. First, the titles and abstracts were screened to identify the target paper. Second, a pilot screening of ten randomly selected papers was conducted to refine the codebook for assessing the papers. The codebook included all the essential elements involved in the performance of MI recommended by Van de Schoot et al. (2012). The following themes were coded in the codebook: 1) data screening: total sample size, group numbers, group size ratio, Item level, analytic software, missing data and methods for dealing with missing data, reliability, descriptive analysis, and data structure; 2) estimation methods and fit indices: estimator used for
model estimation, type of CFA model, number of factors, factor loading size, model fit criteria, and CFA fitting for each group separately; 3) MI testing: levels of full invariance established (configural, metric, scalar, or residual) and levels of partial MI, model fit criteria used for model comparison, design effect (i.e., accounting for clustering effects) considered for MI, and model fit statistics used. Table 2.1 presents the checklist elements included in this review study.

After creating the codebook, the full texts were screened based on the eligibility criteria. The inclusion criteria include 1) used MG-CFA analysis; 2) used empirical data in the analyses and excluded simulated data; 3) conducted data analyses including various levels of invariance testing such as configural, metric, or scalar; and residual invariance; and 4) if full invariance was not achieved, partial invariance was used. In the screening process, the relevant information was entered in alignment with the elements listed in the codebook in an excel file. If uncertainty arose in the screening process, discussions with experts in MI were made to arrive at a final decision.

After reviewing the papers, the percentage of each sub-element under each theme (e.g., the percentage of the studies involving different numbers of groups, the percentage of the studies which established full scalar invariance, the percentage of the studies which considered the design effect) was calculated to examine the frequency of those elements used in these studies. In addition to the variables in the checklist, four new variables were created to indicate each of the four levels of MI (i.e., configural, metric, scalar, and residual). The variables were coded as dichotomous with 0=No (i.e., Non-full invariance) and 1=Yes (i.e., full invariance). Three variables representing the model fit criteria were also created (i.e., $\Delta \chi^2$ only, $\Delta$AFI only, or both $\Delta \chi^2$ and $\Delta$AFI) and coded as
dichotomous with 0=No and 1=Yes. Three separate variables representing group size comparisons were also created, including two groups (0=equal to two groups, 1= having more than two groups), three groups (0=equal to or less than three groups, 1= having more than three groups), and four groups (0=equal to or less than four groups, 1= having more than four groups). Pearson correlation coefficients were calculated to examine the association between each level of full invariance established and model fit criteria or sample characteristics (e.g., group size). All analyses were conducted with SPSS 28.0.

Results

The current study reviewed 63 papers on MI. 91 records were retrieved by searching the Education Resources Information Center Database (ERIC). Five records were excluded from the abstract screening as they were not empirical research. The remaining 86 records were assessed in more detail based on full texts. Of these, 23 were discarded as they did not meet the eligibility criteria. Finally, 63 papers were included in this systematic review. The study selection process is shown in Figure 2.2.

The 63 papers reviewed were published by 51 Journals. The number of papers published by these journals ranged from 1 (e.g., Assessment of Effective Intervention, International Journal of Research and Method in Education) to 18 (e.g., Assessment). The other journals published 5 to 10 MI papers included Personality & Individual Differences (n=5), Measurement and Evaluation in Counseling and Development(n=6), and Studies in Educational Evaluation(n=10). Multiple MI tests were conducted in some of the studies. In these 63 MI papers, 129 MI tests were conducted across different groups with MG-CFA. All the analyses were conducted with the MI test as the unit of analysis rather than paper reviewed.
Data Screening

60.5% of the MI tests analysis was conducted with Mplus software, 18.6% with R, and 8.6% used other software, including AMOS, LISREL, SPSS, and STRATA. 12.4% of the MI tests did not report the software used for MI analysis. All studies except one reported the sample size. The total sample sizes in the studies varied widely, ranging from 95 to 834,428. The median sample size was 1064.

All MI tests reported group numbers. The number of groups involved in MI ranged from 2 to 57. 73.6% of the MI conducted two groups invariance testing, followed by three groups (7.0%), four groups (5.4%), and five groups (1.6%). 12.4% of MI tests involved more than six groups. 17.8% of the MI tests did not report group sizes. The ratios of group sizes also varied across studies, with a minimum ratio of 1:1 and a maximum ratio of 39:1. Most studies had group ratios of 1:1 (67.9%) and 2:1 (18.9%). Most of the groups investigated with MI in educational research were gender (34.4%), culture (20.3%), age (12.5%), grade level (7.8%), race (6.3%), and others (28.8%), including school level, language, informants, SES, platform, clinical status, and treatment condition.

Regarding the data type, 27.9% of the MI tests used items with levels ranging from two to four, 32.0% of the MI tests involved items with five levels, 35.2% used items with six or more levels, 4.0% used both continuous and categorical data in one measurement scale, and 3.1% did not discuss the item scales.

Of the set, 38.8% of studies involved missing data, 16.3% did not have missing data, and 38.3% did not report whether they had missing data. Among 50 MI tests that reported missing data, only 31.0% of studies directly stated their methods for dealing
with missing data. The methods these studies adopted for dealing with missing data included FIML (72.5%), imputation (10.0%), pairwise deletion (7.5%), and others (10.0%) such as pairwise correlation, replacing missing values with series means, pairwise correlation, and default methods for the estimation.

The majority of the studies reported the reliability of the instruments (80.6%). The reported reliability included reliability by Cronbach’s Alpha only (54.8%), omega reliability only (14.4%), Composite reliability only (1.0%), both Alpha and Composite reliability (3.8%), both Alpha and omega (21.2%), both omega and Composite reliability (3.8%), and both omega and Test-retest reliability (1.0%).

Among the studies, 3.1% did not provide information about the data structure. 12.4% of the studies provided some information about the data structure, but more detail is needed to decide whether the data were multilevel. Among the 109 studies that provided clear information about whether data were multilevel, 33.0% were single-level, and 67.0% were multilevel (e.g., students are nested within the classroom). Among the studies with multilevel data, only 13.7% took into account the hierarchical multilevel structure of the data by considering the correlations among observations (e.g., using Type = Complex Mplus option). All the studies explored only the single-level MI while accounting for the nesting nature of the data. No study explored both within-level and between-level MI. Data screening results are presented in Table 2.2

**Estimation Method and Fit Indices**

Most studies (49.6%) used Maximum Likelihood (ML) related estimators (i.e., ML, MLR, FIML), 27.9% used alternative estimators such as WLSMV or ULSMV, and 22.5% of MI tests did not report the estimator used. MI tests with mixed continuous and
categorical data (e.g., items of 2 to 4 levels mixed with items of 5 or more than 5 levels) used FIML (25.0%) or WLSMV (75.0%) as estimators. For the studies using instruments with items of 6 or more levels, 80.5% used ML, MLR, or FIML, and the rest, 19.5%, used WLSMV or ULSMV. For the MI tests with items of 5 levels, 67.9% used ML, MLR, or FIML, and 32.1% used WLSMV or ULSMV. For the MI tests using instruments with items of 2 to 4 levels, 64.0% used WLSMV, and 36.0% used ML, MLR, or FIML.

About 81.4% of the MI tests conducted CFA measurement model testing prior to MI testing. Among these tests, 25.7% of the studies fitted CFA models with each group sample, 52.4% fitted CFA models with the full sample, and 21.9% fitted CFA models with both the group and full sample. Most MI tests used model fit statistics, including chi-square (53.5%), CFI (98.4%), TLI (53.1%), RMSEA (96.1%), and SRMR (58.6%). Few studies (19.3%) used other model fit indexes such as AIC, NFI, and NNFI. 31.0% of the MI tests used all four model fit statistics (i.e., chi-square, CFI, RMSEA, SRMR), and 53.2% used three of the four model fit statistics. The percentage of the studies which used two (15.0%) or only one (0.8%) of the fit statistics was small.

Among the studies which conducted CFA as the baseline model, 82.9% established single-order CFA factor structures, 9.3% identified second-order factor structures, and 7.8% established bi-factor CFA. Among those studies which established single-order CFA structures, the number of factors identified ranged from 1 to 10. The most commonly identified factor numbers are 1 factor (40.2%), two factors (15.0%), three factors (12.1%), and four factors (10.3%). For those identified second-order factor structures, the number of the second-order factors ranged from 1 to 4, and the number of the first-order factors ranged from 2 to 12. Among the studies which conducted CFA
models, 78.1% reported factor loadings. Among these that reported factor loadings, all except one study reported standardized factor loadings. 81.5% of studies reported factor loading sizes equal to or greater than 0.3, and 18.5% reported certain items with factor loadings below 0.3. MG-CFA model fitting results are presented in Table 2.3.

**Measurement Invariance Testing**

The model fit statistics used to evaluate MI were categorized into three types: only the chi-square differences (Δχ²), only alternative fit indices differences (ΔAFI; e.g., ΔCFI, ΔREMSEA, and ΔSRMR), and both Δχ² and ΔAFI. As 37 studies reported chi-square values without considering the significance of the value, these studies were treated as not using chi-square. 11.10% of the studies reported only Δχ², 73.8% reported ΔAFI, and 15.1% reported both Δχ² and ΔAFI (Table 2.4).

Different cutoff values for fit statistics were used among MI tests using ΔAFI fit statistics for evaluating MI. Regarding ΔCFI, 83.5% used a cutoff value of 0.01, 5.8% used 0.005, and 2.9% used 0.02. Also, 7.8% used 0.01 for metric invariance and 0.02 for scalar invariance. Different ΔREMSEA cutoff values were used, including 0.015(72.4%), 0.01(17.2%), 0.01 for metric invariance, 0.03 for scalar invariance (9.2%), and 0.015 for metric and 0.03 for scalar invariance (1.1%). Regarding ΔSRMR, 51.4% of the MI tests used the cutoff values of 0.03, 11.4% used 0.015, 22.9% used 0.01 for metric invariance and 0.03 for scalar invariance, and 14.3% used 0.015 for metric invariance and 0.03 for scalar invariance. Overall, ΔCFI≤0.010, ΔRMSEA≤0.015, and ΔSRMR≤0.30 were the most frequently used criteria for evaluating MI (Table 2.5)

Table 2.6 presents the percentage of the MI tests that established each of the four levels. All studies except one clearly stated the levels they tested. Among these MI tests
which reported the invariance level they tested, all invariance tests included a configural invariance test, and 96.9% of the tests included metric invariance. 82.0% tested scalar invariance, but only 20.3% included a test of residual invariance. Most comparisons established full configural (97.7%) and metric invariance (91.9%). The percentage of the full scalar and residual invariance established was 84.8% and 88.0%, respectively. The percentages of the partial metric (8.1%), scalar (15.2%), and residual (12.0%) MI established were relatively small. No invariance was established only from configural invariance tests (2.3%). Overall, 19.5% of the MI tests reported partial invariance for one of the three steps (metric, scalar, and residual). Among the MI tests with full scalar invariance (n=89), 39.3% examined the latent mean differences. Among the studies with partial scalar invariance (n=16), 43.8% tested the latent mean differences across groups by freeing 1 to 8 parameters.

Table 2.7 displays the correlation between levels of full invariance identified and model fit statistics and sample characteristics. The sample size did not have a significant association with any of the MI levels. The criteria used for model comparison to decide the MI were not associated with the level of invariance established. Group number (\( \leq 2 \) vs>2) did not show a significant association with the identification of the full MI identification. Group number (\( \leq 3 \) vs>3) shows a significant association with full configural invariance. MI tests involving three groups or fewer than three groups are more likely to establish the configural invariance (\( r = -.184 \)) relative to those involving more than three groups. Group number (\( \leq 4 \) vs>4) shows a significant association with the identification of full configural invariance and residual invariance. MI tests involving four groups or fewer than four groups are more likely to establish
full configural invariance \( (r=-.234) \) or full residual invariance \( (r=-.553) \).

**Discussion**

The number of MI practices conducted in education has increased in the last decade. MI is typically tested with MG-CFA in the SEM framework. This study systematically reviewed current practices for testing MI with the MG-CFA approach in education. The study raises concerns regarding what is expected to report in MI-related studies and what is needed for future research reports. More specifically, the following aspects of MI practices were reviewed, including data screening (e.g., sample size, group number, group size ratio, item level, analytic software, data missing, reliability, descriptive analysis, and data structure), common measurement model fit (e.g., estimator used, type of CFA model, factor number, factor loading size, model fit criteria, separate group model fitting), and measurement invariance testing (e.g., levels of full invariance, levels of partial invariance, model fit statistics for model comparison, and multilevel data structure consideration).

**Data Screening**

Van de Schoot et al. (2012) suggested that data be properly screened before testing MI. The sample size is of primary concern when researchers design a study (Tabachnick et al., 2013). Sample sizes involved in the studies reviewed range from 95 to 834,428, with a mean sample size of 56,970. Sample size affects the power of the study and whether there are a sufficient number of cases for the model to converge for a proper solution and accurate parameter estimates. A sample size range from 300 to 460 cases is recommended for structural equation modeling (SEM) while considering the number of indicators and factors, the magnitude of factor loadings, path coefficients, and amount of
missing data (Wolf et al., 2013). However, 8.7% of the studies on MI sampled less than 300 participants. Within the SEM framework, a small sample size causes estimation convergence failure, improper solution, inaccurate parameter estimates and model fit statistics (Wang & Wang, 2019). A ratio of cases to free parameters (n: p) was recommended as a CFA/SEM rule of thumb for minimum sample size. The commonly used (n:p) ratio is 10:1 to 20:1 (Kline, 2015).

Roughly 3.1% of the MI reviewed did not discuss the item scales. However, the model estimation method, which depends on the data type (e.g., continuous or categorical), may affect the model fit indices (Xia & Yang, 2019). For example, multivariate normal distribution should be met under the ML analysis. The reporting of the data type is recommended for deciding whether the appropriate model estimation method was chosen for model fitting.

Regarding missing data reporting, 38.3% of studies did not report whether they had missing data. Ignoring missing data leads to inaccurate results. For example, when nonnormality and missingness increased, the ML chi-square test became vulnerable and the model rejection rates increased (Savalei, 2008). Missing data could impact MI (Chen et al., 2020; Selvi et al., 2020; Tsai & Yang, 2012). Therefore, discussing missing patterns and dealing methods is necessary if missing data are present. Among the studies with missing data, 30.1% of the studies stated directly the methods they used for dealing with missing data. The most frequently used method is the full information maximum likelihood estimation method (FIML), followed by imputation, pairwise deletion, and other methods. Continuous full information maximum likelihood estimation method (FIML), continuous robust FIM (rFIML), FIML with probit links (pFIML), FIML with
logit links (IFIML), or the mean and variance adjusted weight least squared estimation method (WLSMV) combined with pairwise deletion (WLSMV_PD) were suggested for dealing with missing ordinal data in MI testing (see details in Chen et al., 2020). Studies with around 5% of missing data in large sample sizes can use regression imputation, expectation-maximization, and multiple imputation methods to deal with missing data (Selvi et al., 2020).

Reliability refers to an instrument’s consistency (Barry, 2014). Specifically, internal consistency reliability examines the uniformity of results across items of a scale. About 19.4% of the studies did not report the reliability of the instruments. Failure to report the reliability may lead to uncertainty regarding whether the instruments employed in the study were able to produce consistent scores. Therefore, researchers and practitioners should assess and report a scale’s reliability and validity when an instrument is administered to ensure the integrity of data collected from the scale (Barry et al., 2014). Of the coefficients, Cronbach’s Alpha coefficients (α) and McDonald’s Omega coefficient (ω) are two common ways of measuring reliability (Ravinder & Saraswathi, 2020). Although Cronbach’s Alpha coefficient (α) is frequently reported, the assumptions of tau-equivalent (i.e., equal factor loadings of all items in a factorial model) are seldom met in practice (Revelle & Zinbarg, 2009; Shevlin et al., 2000), and it has the disadvantage of underestimating the true reliability when data are multidimensional (Osburn, 2000). Therefore, the alpha coefficient (α) is not recommended for estimating the reliability of a multidimensional scale (Gignac, 2014). Omega (ω) is less likely to be misinterpreted because of the transparent relationship with the factor models they are based on (Green & Yang, 2015); however, ω can be biased if the underlying model is
misspecified. Further, the index requires large overall samples (Green & Yang, 2015). Therefore, it is suggested that researchers report both coefficient alpha and omega to make reliability less likely to be misinterpreted (Green & Yang, 2015).

The data structure was also examined, and 67.0% were multilevel. However, only 13.7% of the studies accounted for the multilevel structure of the data. When the data structure is hierarchical, the assumption of independent observations is not met for single-level analysis. In this case, researchers should deal with the correlations among observations (e.g., multilevel modeling) (Raudenbush & Bryk, 2002). If researchers ignore the nested data structure, especially when ICC values are large, the chi-square global fit index is inflated, the unstandardized parameter estimates are overestimated, and standard errors are underestimated (Julian, 2001; Pornprasertmanit et al., 2014). In MI tests, the invariant model is more likely to be rejected and detected as being noninvariant if the dependency of the data is not taken into account for the analysis (Kim et al., 2012). Design-based and model-based procedures are suggested for accounting for nested data (Stapleton, 2013). The design-based approach focused the analysis on individual levels while considering the nested structure of the data as part of the design process (Julian, 2001). All the studies considering nested data structure tested only within-group invariance with a design-based approach. The model-based procedure allows researchers to examine within-cluster and between-cluster relations simultaneously (Stapleton, 2013). Researchers interested in both within-level and between-level invariance can use multi-group multilevel-confirmatory factor analysis (MGM-CFA) for detecting two-level invariance.
Estimation Methods and Fit Indices

Models are misspecified to some degree due to various factors in the application of SEM (MacCallum, 2003). One of the factors is the choice of estimation method, which may affect the model fit indices (Brosseau-Liard et al., 2012; Xia & Yang, 2019). However, among the MI studies reviewed, 22.5% did not report the estimator used. The use of estimators was inconsistent. Both WLSMV-related and ML-related estimators were used regardless of the data type (i.e., whether they are continuous or categorical). For example, among the studies involving items of two to four levels, 64.0% used WLSMV, and 36.0% used ML, MLR, or FIML. The choice of the estimator is, to a large extent, determined by the nature of the data. Specifically, Maximum Likelihood (ML) is preferred for continuous data under the assumptions of independence and multivariate normality (Kline, 2015). When ML is used with ordinal data without meeting the assumptions, standard errors will not be consistent (Lubke & Muthén, 2004). Weighted Least Square estimators (i.e., DWLS, WLSM, WLS, or WLSMV) are recommended to account for the categorical nature of data (Finney & DiStefano, 2002), especially when sample sizes are larger than 500 (Bandalos, 2008). Among these WLS-based estimators for use with categorical data, DWLS and ULS produced similar parameter estimates, standard errors, and mean-and-variance-adjusted goodness-of-fit statistics with moderate sample sizes (e.g., N=500 or 600). However, ULS performed slightly better than DWLS with small sample sizes (e.g., N=200) (Forero et al., 2009; Savalei & Rhemtulla, 2013). WLSMV and ULSMV outperformed WLSM and ULSM in small samples (Savalei & Rhemtulla, 2013). Researchers need to choose appropriate estimators based on the research settings, such as data types, data distribution, and sample sizes.
Within the MI framework, the choice of estimator affects the identification of the MI levels. For example, when testing metric and scalar invariance, Type I error rates produced by $\Delta \chi^2$ varied substantially across estimators (ML, MLR, and WLSMV) with symmetric and asymmetric data when models were misspecified. However, the tested estimator performed similarly regardless of the estimator when models were correctly specified (Sass et al., 2014). It is suggested that researchers use various estimators to determine the stability of conclusions and report the model results that are most justifiable statistically and theoretically (Sass, et al., 2014). Also, researchers should select estimation methods based on the data type and check whether the assumptions of using the estimator are met.

MI should start with specifying a CFA that reflects how the construct is theoretically operationalized by fitting CFA models for each group separately to ensure adequate factorial validity (Sass, 2011; Van de Schoot et al., 2012). However, only 38.7% of the studies established a common measurement model for each group. Therefore, future researchers should assess the model fit for each group separately to examine the factorial validity (Sass et al., 2014). Chi-square (53.5%), CFI (98.4%), TLI (53.1%), RMSEA (96.1%), and SRMR (58.6%) were the four most frequently used fit statistics in the current study. The chi-square test, known as the exact-fit test, assesses the discrepancy between the model implied covariances and the observed sample covariances. Although the chi-square test is sensitive to large sample sizes and or model complexity, it still has its benefits and needs to be reported in the same way one would report the SEM model fit based on other statistical tests (Barrett, 2007). Approximate fit statistics (i.e., CFI, RMSEA, and SRMR) can adjust for sample size and model
complexity. However, the guidelines regarding what constitutes good model fit are subjective and vary across different models (Hu & Bentler, 1999; Marsh et al., 2004). Therefore, it is suggested that a minimum set of fit statistics should be reported, including Chi-square with its degree of freedom and $p$-value, RMSEA, CFI, and SRMR (Kline, 2015). The current study found that 31.0% of the studies reported all these four fit statistics. Future research should consider reporting at least four of these fit statistics. The current study found that 21.9% of the studies did not report the local model fit indices, such as factor loadings and standardized residuals. However, as global model fit might be affected by the magnitude of the factor loadings and the number of items (Greiff & Heene, 2017; McNeish et al., 2018), it is suggested that local model fit be reported as well.

**Measurement Invariance Testing**

This study found that most invariance tests were conducted at less strict levels, and researchers mainly focused on the full invariance test. For example, all of the tests were conducted at the configural level, 96.9% of tests at the metric level, and 82.0% of tests at the scalar level, but only 20.3% of tests were conducted at the residual level. The small percentage of residual MI tests might be due to the findings that scalar invariance is adequate for latent mean comparison across groups (Steenkamp & Baumgartner, 1998).

The partial invariance test allows researchers to investigate model invariance after freeing the noninvariant parameters (Byrne et al., 1999). It indicates that only a subset of the subgroup parameters (i.e., some factor loadings and item intercepts) is the same across groups (Schmitt & Kuljanin, 2008). Among the studies which did not establish full invariance, only 8.1%, 15.2%, and 12.0% of the tests conducted metric, scalar, or residual
partial invariance, respectively, by freeing parameters. Partial invariances impact comparison of observed composite differences (Steinmetz, 2013) and cross-group comparisons of the differences in latent means, latent variances, or structural relations with other constructs (Cheung & Rensvold, 2002; Schmitt & Kuljanin, 2008). Researchers need to consider a partial invariance test when a certain level of invariance is not supported by the full invariance test, as fitting partially invariant models yielded more accurate estimates of the parameters of interest (Shi et al., 2019).

Regarding model evaluation, the choice of model fit indices and evaluating criteria for MG-CFA needs to be clearly stated. However, 2.8% of the studies did not report model fit indices for model comparison. The choice of the fit measures used to support invariance was inconsistent. Approximately 11.1% of the studies used only Δχ², 73.8% reported only ΔAFI, and 15.1% reported both Δχ² and ΔAFI. A non-significant Chi-square difference would provide some statistical evidence for the compatibility between the two invariance models tested, but this test is sensitive to large sample sizes, model complexity, nonnormal data, and model misspecification (Asparouhov & Muthén, 2006; Chen, 2007; Cheung & Rensvold, 2002; French & Finch, 2008; Mead & Bauer, 2007; Yuan & Bentler, 2004). Alternative fit indices (ΔAFI) reflected the changes in fit between a more restrictive model and a less restrictive model (Sass et al., 2014). Relative to Δχ², ΔAFI were less affected by sample size, the number of factors, or the number of items (Chen, 2007; Cheung & Rensvold, 2002; Mead et al., 2008). However, ΔAFI may not perform well with misspecified models with ordinal data (Sass et al., 2014). Besides, ΔAFI was also affected by the proportion of invariant indicators, the pattern of invariance, and equality of group sizes (Chen, 2007). Even among ΔAFI, ΔCFI and
ΔRMSEA were equally sensitive to factor loadings and intercept differences, while ΔSRMR was more sensitive to noninvariant factor loadings than intercept (Chen, 2007).

Given the limitations of the Δχ² and ΔAFI, researchers need to consider the statistical significance of the Δχ² after a Bonferroni adjustment, change in approximated fit statistics, and the magnitude of differences between the parameter estimates (Sass et al., 2014). The choice of fit measures was not associated with identifying any MI level, meaning that researchers can choose multiple fit measures without considering the MI levels they are to test.

Among the MI tests using ΔAFI for evaluating MI, the criteria used were inconsistent. For example, ΔCFI cutoff values included 0.01, 0.02, and 0.005. Even for the same fit measures, some MI tests used different criteria for detecting different invariance levels. For example, some MI tests used ΔCFI cutoff value of 0.01 for detecting metric invariance and 0.02 for scalar invariance. ΔCFI≤.010, ΔRMSEA≤.015, and ΔSRMR≤.30 were the three most commonly used ΔAFI criteria for evaluating MI. Previous research found that model size (Cao & Liang, 2022; Cheung & Rensvold, 2002; French & Finch, 2008, Meade et al., 2008) and group size ratio (Chen, 2007; Yoon & Lai, 2008) affect the sensitivity of the fit measures to MI testing. Researchers might need to consider these factors while choosing the fit measures for detecting MI.

This study explored whether sample size and the number of groups being compared affect the establishment of certain levels of MI. The sample size did not show a significant impact on achieving any level of MI. This might be due to the less frequent use of Δχ² (11.1%), which is sensitive to sample size, and more researchers chose ΔAFI only as criteria for MI establishment (73.8%). The number of groups involved in MI tests
was associated with establishing a certain level of MI. Comparing two vs. more than two groups was not associated with establishing any MI level. MI tests involving more than three groups were less likely to establish configural invariance than those with three or fewer groups. Comparing four vs. more than four groups was associated with establishing full configural and residual invariance. Relative to studies with more than four groups, studies involving four or fewer groups are more likely to establish full configural invariance or full residual invariance.

**Conclusion**

Measurement invariance indicates that the same latent construct is measured similarly across different measurement conditions. The establishment of MI ensures that the cross-condition differences in observed variables reflect true changes in latent constructs rather than the changes in the psychometric properties of the measures. MI is a prerequisite for making meaningful comparisons across different measurement conditions (Meredith & Teresi, 2006). In the educational context, measuring instruments were often developed to assess constructs that are not directly observable. In this scenario, MI tests are especially important for ensuring the validity of the instrument. This study conducted a systematic review of the MI practices in education by focusing on the PRISMA protocol items developed by Liberati et al. (2009). The results of the current study showcased the limitations of the MI practices conducted by researchers in the field of education.

Regarding data screening, around one-fourth of the MI tests did not report the group number involved for comparison or reliability of the measures. Approximately 38.3% of the tests did not report missing data. 86.3% of the MI tests did not consider the
multi-level nature of the data structure. With respect to common measurement model fitting, around one-fourth of the MI tests did not report the estimator used for model estimation. Less than half of the tests tested CFA with a single group sample. Around 31.0% of MI tests did not report at least four model fit statistics, including chi-square, CFI, RMSEA, and SRMR. Regarding MI testing, the current study found that only 15.1% of the tests used both Δ $\chi^2$ and Δ AFI for model comparisons. Among the MI tests using Δ AFI for evaluating MI, the choice of the fit measures criteria was inconsistent. This study also found that the three-group comparison is more likely to achieve configural invariance than studies comparing more than three groups. Studies comparing four groups are more likely to establish configural invariance and residual invariance relative to studies with more than four groups.

This study reviewed the major elements to be considered when conducting MI analyses and offered suggestions regarding MI reporting. However, this study touched only on many of the issues without providing an in-depth analysis of all the elements to be considered across different invariance models. For example, group size and power were not discussed, and partial invariance technics were discussed less. In summary, researchers may consider conducting MI tests by following the recommendation mentioned in this study.

References


https://doi.org/10.1080/10705510801922340

https://doi.org/10.1016/j.paid.2006.09.018

https://doi.org/10.1177/1090198113483139

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


https://doi.org/10.1037/0033-2909.88.3.588

https://doi.org/10.1097/01.mlr.0000245143.08679.cc

Brosseau-Liard, P. E., Savalei, V., & Li, L. (2012). An investigation of the sample performance of two nonnormality corrections for RMSEA. *Multivariate behavioral research, 47*(6), 904-  
https://doi.org/10.1080/00273171.2012.715252


https://doi.org/10.1177/0022022199030005001

https://doi.org/10.1080/10705511.2022.2056893


[https://doi.org/10.1097/01.mlr.0000245438.73837.89](https://doi.org/10.1097/01.mlr.0000245438.73837.89)


[https://doi.org/10.1037/1082-989x.5.3.343](https://doi.org/10.1037/1082-989x.5.3.343)

[https://doi.org/10.1177/1094428103259554](https://doi.org/10.1177/1094428103259554)

[https://doi.org/10.1080/00273171.2014.933762](https://doi.org/10.1080/00273171.2014.933762)

[https://doi.org/10.1016/j.dr.2016.06.004](https://doi.org/10.1016/j.dr.2016.06.004)


[https://doi.org/10.1037/0033-2909.114.3.552](https://doi.org/10.1037/0033-2909.114.3.552)

[https://doi.org/10.1177/001316449805806010](https://doi.org/10.1177/001316449805806010)


[https://doi.org/10.1177/1073191117711020](https://doi.org/10.1177/1073191117711020)


[https://doi.org/10.1080/10705511.2017.1387859](https://doi.org/10.1080/10705511.2017.1387859)

[https://doi.org/10.1177/0013164404264853](https://doi.org/10.1177/0013164404264853)
Table 2.1 Checklist for MG-CFA review

<table>
<thead>
<tr>
<th>Themes</th>
<th>Sub-themes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Screening</td>
<td>Sample Size</td>
</tr>
<tr>
<td></td>
<td>Group number</td>
</tr>
<tr>
<td></td>
<td>Group size ratio</td>
</tr>
<tr>
<td></td>
<td>Item level</td>
</tr>
<tr>
<td></td>
<td>Analytic software</td>
</tr>
<tr>
<td></td>
<td>Data missing/missing data dealing methods</td>
</tr>
<tr>
<td></td>
<td>Reliability</td>
</tr>
<tr>
<td></td>
<td>Descriptive analysis</td>
</tr>
<tr>
<td></td>
<td>Data structure</td>
</tr>
<tr>
<td>Estimation Methods and Fit Indices</td>
<td>Estimator</td>
</tr>
<tr>
<td></td>
<td>Type of CFA model</td>
</tr>
<tr>
<td></td>
<td>Factor number</td>
</tr>
<tr>
<td></td>
<td>Factor loading size</td>
</tr>
<tr>
<td></td>
<td>Model fit criteria</td>
</tr>
<tr>
<td></td>
<td>Single group CFA fitting</td>
</tr>
<tr>
<td></td>
<td>Levels of full invariance</td>
</tr>
<tr>
<td></td>
<td>Levels of partial invariance/Numbers of parameters freed</td>
</tr>
<tr>
<td>Measurement Invariance Testing</td>
<td>Model fit statistics used for model comparison</td>
</tr>
<tr>
<td></td>
<td>Design effect considered</td>
</tr>
<tr>
<td></td>
<td>Model fit statistics</td>
</tr>
<tr>
<td></td>
<td>Fit statistics criteria</td>
</tr>
</tbody>
</table>

Note. CFA= confirmatory factor analysis
Table 2.2 Data screening results from MG-CFA review articles

<table>
<thead>
<tr>
<th>Data Screening Elements</th>
<th>Sub-elements</th>
<th>N</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Software</td>
<td>Mplus</td>
<td>78</td>
<td>60.5%</td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>24</td>
<td>18.6%</td>
</tr>
<tr>
<td></td>
<td>Others</td>
<td>11</td>
<td>8.6%</td>
</tr>
<tr>
<td></td>
<td>Not reported</td>
<td>16</td>
<td>12.4%</td>
</tr>
<tr>
<td>Group Number</td>
<td>2 groups</td>
<td>95</td>
<td>73.6%</td>
</tr>
<tr>
<td></td>
<td>3 groups</td>
<td>9</td>
<td>7.0%</td>
</tr>
<tr>
<td></td>
<td>4 groups</td>
<td>7</td>
<td>5.4%</td>
</tr>
<tr>
<td></td>
<td>5 groups</td>
<td>2</td>
<td>1.6%</td>
</tr>
<tr>
<td></td>
<td>6 or above 6 groups</td>
<td>16</td>
<td>12.4%</td>
</tr>
<tr>
<td>Group Size Ratio</td>
<td>1:1</td>
<td>72</td>
<td>67.9%</td>
</tr>
<tr>
<td></td>
<td>1:2</td>
<td>20</td>
<td>18.9%</td>
</tr>
<tr>
<td></td>
<td>1:3 or above 3</td>
<td>14</td>
<td>13.0%</td>
</tr>
<tr>
<td></td>
<td>Not reported</td>
<td>23</td>
<td>17.8%</td>
</tr>
<tr>
<td>Data Type</td>
<td>Items with 2-4 levels</td>
<td>36</td>
<td>28.8%</td>
</tr>
<tr>
<td></td>
<td>Items with 5 levels</td>
<td>40</td>
<td>32.0%</td>
</tr>
<tr>
<td></td>
<td>Items with levels ≥ 6</td>
<td>44</td>
<td>35.2%</td>
</tr>
<tr>
<td></td>
<td>Mixed</td>
<td>5</td>
<td>4.0%</td>
</tr>
<tr>
<td></td>
<td>Not reported</td>
<td>5</td>
<td>3.1%</td>
</tr>
<tr>
<td>Missing</td>
<td>With missing data</td>
<td>50</td>
<td>38.8%</td>
</tr>
<tr>
<td></td>
<td>Without missing data</td>
<td>21</td>
<td>16.3%</td>
</tr>
<tr>
<td></td>
<td>Not reported</td>
<td>58</td>
<td>38.3%</td>
</tr>
<tr>
<td>Missing Data Dealing Method</td>
<td>FIML</td>
<td>29</td>
<td>72.5%</td>
</tr>
<tr>
<td></td>
<td>Imputation</td>
<td>4</td>
<td>10.0%</td>
</tr>
<tr>
<td></td>
<td>Pairwise deletion</td>
<td>3</td>
<td>7.5%</td>
</tr>
<tr>
<td></td>
<td>Others</td>
<td>4</td>
<td>10.0%</td>
</tr>
<tr>
<td>Reliability Reported</td>
<td>Yes</td>
<td>104</td>
<td>80.6%</td>
</tr>
<tr>
<td>Reliability Coefficients</td>
<td>No</td>
<td>25</td>
<td>19.4%</td>
</tr>
<tr>
<td>-----------------------------------------</td>
<td>------</td>
<td>------</td>
<td>-------</td>
</tr>
<tr>
<td>Alpha only</td>
<td>57</td>
<td>54.8%</td>
<td></td>
</tr>
<tr>
<td>Omega only</td>
<td>15</td>
<td>14.4%</td>
<td></td>
</tr>
<tr>
<td>Composite only</td>
<td>1</td>
<td>1.0%</td>
<td></td>
</tr>
<tr>
<td>Both alpha and composite</td>
<td>4</td>
<td>3.8%</td>
<td></td>
</tr>
<tr>
<td>Both alpha and Omega</td>
<td>22</td>
<td>21.2%</td>
<td></td>
</tr>
<tr>
<td>Both Omega and Composite</td>
<td>4</td>
<td>3.8%</td>
<td></td>
</tr>
<tr>
<td>Both Omega and Test-retest</td>
<td>1</td>
<td>1.0%</td>
<td></td>
</tr>
<tr>
<td>Descriptive Analysis</td>
<td>Yes</td>
<td>85</td>
<td>65.9%</td>
</tr>
<tr>
<td>No</td>
<td>44</td>
<td>34.1%</td>
<td></td>
</tr>
<tr>
<td>Data Structure</td>
<td>Multilevel</td>
<td>73</td>
<td>67.0%</td>
</tr>
<tr>
<td>Non-multilevel</td>
<td>36</td>
<td>33.0%</td>
<td></td>
</tr>
<tr>
<td>Multi-level data structure in analysis</td>
<td>Considered</td>
<td>10</td>
<td>13.7%</td>
</tr>
<tr>
<td>Not considered</td>
<td>73</td>
<td>86.3%</td>
<td></td>
</tr>
</tbody>
</table>

Note. Software/Others: include AMOS, LISREL, SPSS, and STRATA; Data Type/Mix: a scale with items of 2-4 levels and items of equal to or more than 5 levels; Missing data dealing method/others: pairwise correlation, replacing missing values with series means, pairwise correlation, and default methods for the estimation; FIML=full information maximum likelihood;

Table 2.3 Estimation and fit indices used in MG-CFA review articles

<table>
<thead>
<tr>
<th>CFA Model Fitting themes</th>
<th>Sub-elements</th>
<th>N</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimator</td>
<td>ML/MLR/ FIML</td>
<td>64</td>
<td>49.6%</td>
</tr>
<tr>
<td></td>
<td>WLSMV/ULSMV</td>
<td>36</td>
<td>27.9%</td>
</tr>
<tr>
<td></td>
<td>Not reported</td>
<td>29</td>
<td>22.5%</td>
</tr>
<tr>
<td>Estimator used for different types</td>
<td>ML/MLR/ FIML (Items of 6 or more than 6 levels)</td>
<td>33</td>
<td>80.5%</td>
</tr>
<tr>
<td>of data</td>
<td>WLSMV/ULSMV (Items of 6 or more than 6 levels)</td>
<td>8</td>
<td>19.5%</td>
</tr>
<tr>
<td></td>
<td>ML/MLR/ FIML (Items of 5 levels)</td>
<td>19</td>
<td>67.9%</td>
</tr>
<tr>
<td></td>
<td>WLSMV/ULSMV (Items of 5 levels)</td>
<td>9</td>
<td>32.1%</td>
</tr>
<tr>
<td></td>
<td>ML/MLR/ FIML (Items of 2 to 4 levels)</td>
<td>9</td>
<td>36.0%</td>
</tr>
<tr>
<td></td>
<td>WLSMV/ULSMV (Items of 2 to 4 levels)</td>
<td>16</td>
<td>64.0%</td>
</tr>
<tr>
<td>FIML (Items of 2 to 4 levels and items of 5 or more than 5 levels mixed)</td>
<td>1</td>
<td>25.0%</td>
<td></td>
</tr>
<tr>
<td>WLSMV/ULSMV (Items of 2 to 4 levels and items of 5 or more than 5 levels mixed)</td>
<td>3</td>
<td>75.0%</td>
<td></td>
</tr>
<tr>
<td><strong>CFA model specifying</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CFA with single group sample</td>
<td>27</td>
<td>25.7%</td>
<td></td>
</tr>
<tr>
<td>CFA with full sample</td>
<td>55</td>
<td>52.4%</td>
<td></td>
</tr>
<tr>
<td>CFA with both single groups and full sample</td>
<td>23</td>
<td>21.9%</td>
<td></td>
</tr>
<tr>
<td><strong>Model Fit statistics</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chi-square</td>
<td>69</td>
<td>53.5%</td>
<td></td>
</tr>
<tr>
<td>RMSEA</td>
<td>123</td>
<td>96.1%</td>
<td></td>
</tr>
<tr>
<td>CFI</td>
<td>126</td>
<td>98.4%</td>
<td></td>
</tr>
<tr>
<td>TLI</td>
<td>68</td>
<td>53.1%</td>
<td></td>
</tr>
<tr>
<td>SRMR</td>
<td>75</td>
<td>58.6%</td>
<td></td>
</tr>
<tr>
<td>AIC</td>
<td>3</td>
<td>2.3%</td>
<td></td>
</tr>
<tr>
<td>BIC</td>
<td>0</td>
<td>0.0%</td>
<td></td>
</tr>
<tr>
<td>NFI</td>
<td>5</td>
<td>3.9%</td>
<td></td>
</tr>
<tr>
<td>NNFI</td>
<td>7</td>
<td>5.5%</td>
<td></td>
</tr>
<tr>
<td>IFI</td>
<td>2</td>
<td>1.6%</td>
<td></td>
</tr>
<tr>
<td>GFI</td>
<td>4</td>
<td>3.1%</td>
<td></td>
</tr>
<tr>
<td>AGFI</td>
<td>3</td>
<td>2.3%</td>
<td></td>
</tr>
<tr>
<td>ECVI</td>
<td>2</td>
<td>1.6%</td>
<td></td>
</tr>
<tr>
<td>4 of the 4 fit statistics (Chi-square, CFI, RMSEA, SRMR)</td>
<td>39</td>
<td>31.0%</td>
<td></td>
</tr>
<tr>
<td>3 of the 4 fit statistics (Chi-square, CFI, RMSEA, SRMR)</td>
<td>67</td>
<td>53.2%</td>
<td></td>
</tr>
<tr>
<td>2 of the 4 fit statistics (Chi-square, CFI, RMSEA, SRMR)</td>
<td>19</td>
<td>15.1%</td>
<td></td>
</tr>
<tr>
<td>1 of the 4 fit statistics (Chi-square, CFI, RMSEA, SRMR)</td>
<td>1</td>
<td>0.8%</td>
<td></td>
</tr>
<tr>
<td><strong>Type of CFA models</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single-order CFA</td>
<td>107</td>
<td>82.9%</td>
<td></td>
</tr>
<tr>
<td>Bi-factor CFA</td>
<td>10</td>
<td>7.8%</td>
<td></td>
</tr>
<tr>
<td>Second-order CFA</td>
<td>12</td>
<td>9.3%</td>
<td></td>
</tr>
<tr>
<td><strong>Factor number for Single-order CFA</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 factor</td>
<td>43</td>
<td>40.2%</td>
<td></td>
</tr>
<tr>
<td>2 factors</td>
<td>16</td>
<td>15.0%</td>
<td></td>
</tr>
<tr>
<td>3 factors</td>
<td>13</td>
<td>12.1%</td>
<td></td>
</tr>
<tr>
<td>4 factors</td>
<td>11</td>
<td>10.3%</td>
<td></td>
</tr>
</tbody>
</table>
5 factors 8 7.5%
6 factors 7 6.5%
7 factors 7 6.5%
10 factors 2 1.9%

<table>
<thead>
<tr>
<th>Loading Size reported</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>82</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>78.1%</td>
<td>21.9%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Factor loadings</th>
<th>≥0.30</th>
<th>&lt;0.3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>66</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>81.5%</td>
<td>18.5%</td>
</tr>
</tbody>
</table>

Note. ML=maximum likelihood; MLR=Maximum likelihood with robust standard errors; FLML=full information maximum likelihood; WLSMV=weighted least square mean and variance adjusted; ULSMV=unweighted least square mean and variance adjusted; CFA=confirmatory factor analysis; RMSEA=root mean square error of approximation; CFI=comparative fit index; TLI=Tucker-Lewis index; SRMR=standardized root mean squared residual; AIC=Akaike information criterion; BIC=Bayesian information criterion; NFI=Normed fit index; NNFI=Non-normed fit index; IFI=Incremental fit index; GFI=Goodness of fit; AGFI=Adjusted goodness of fit; ECVI=Expected cross validation index.

Table 2.4 Model fit indices used in MG-CFA review articles

<table>
<thead>
<tr>
<th>Model fit statistics</th>
<th>N</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Δχ² only</td>
<td>14</td>
<td>11.1%</td>
</tr>
<tr>
<td>Δ AFI only</td>
<td>93</td>
<td>73.8%</td>
</tr>
<tr>
<td>Both Δχ² and Δ AFI</td>
<td>19</td>
<td>15.1%</td>
</tr>
</tbody>
</table>

Note. AFI=alternative fit indices
Table 2.5 Fit statistics criteria used in MG-CFA review articles

<table>
<thead>
<tr>
<th>Cutoff criteria</th>
<th>$\Delta$CFI</th>
<th>$\Delta$RMSEA</th>
<th>$\Delta$SRMR</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>83.5%</td>
<td>17.2%</td>
<td>-----</td>
</tr>
<tr>
<td>0.02</td>
<td>2.9%</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>0.005</td>
<td>5.8%</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>0.015</td>
<td>-----</td>
<td>72.4%</td>
<td>11.4%</td>
</tr>
<tr>
<td>0.03</td>
<td>-----</td>
<td>-----</td>
<td>51.4%</td>
</tr>
<tr>
<td>0.01 for metric and 0.02 for scalar invariance testing</td>
<td>7.8%</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>0.015 for metric and 0.03 for scalar invariance testing</td>
<td>-----</td>
<td>1.1%</td>
<td>-----</td>
</tr>
<tr>
<td>0.010 for metric and 0.03 for scalar invariance testing</td>
<td>-----</td>
<td>9.2%</td>
<td>22.9%</td>
</tr>
</tbody>
</table>

Note. $\Delta$=difference, CFI=comparative fit index; RMSEA=root mean square error of approximation; SRMR=standardized root mean squared residual;

Table 2.6 Percentages of MI levels established in MG-CFA review articles

<table>
<thead>
<tr>
<th></th>
<th>Configural invariance (n=127; 100%)</th>
<th>Metric invariance (n=124; 96.9%)</th>
<th>Scalar invariance (n=105; 82.0%)</th>
<th>Residual invariance (n=26; 20.3%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Invariance</td>
<td>2.3%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Partial Invariance</td>
<td>8.1%</td>
<td>15.2%</td>
<td>12.0%</td>
<td></td>
</tr>
<tr>
<td>Full invariance</td>
<td>97.7%</td>
<td>91.9%</td>
<td>84.8%</td>
<td>88.0%</td>
</tr>
</tbody>
</table>
Table 2.7 Pearson correlations of model fit criteria and sample characteristics with levels of invariance

<table>
<thead>
<tr>
<th></th>
<th>Configural invariance</th>
<th>Metric invariance</th>
<th>Scalar invariance</th>
<th>Residual invariance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Δ χ² only</td>
<td>-.118</td>
<td>-.101</td>
<td>.026</td>
<td>-.345</td>
</tr>
<tr>
<td>ΔAFI only</td>
<td>.148</td>
<td>.036</td>
<td>.036</td>
<td>.369</td>
</tr>
<tr>
<td>Both Δ χ² and Δ AFI</td>
<td>-.079</td>
<td>.041</td>
<td>-.063</td>
<td>-.175</td>
</tr>
<tr>
<td>Total sample size</td>
<td>-.153</td>
<td>.091</td>
<td>.100</td>
<td>.157</td>
</tr>
<tr>
<td>Group number (≤2 vs&gt;2)</td>
<td>-.141</td>
<td>.039</td>
<td>.105</td>
<td>-.345</td>
</tr>
<tr>
<td>Group number (≤3 vs&gt;3)</td>
<td>-.184*</td>
<td>.065</td>
<td>.106</td>
<td>---</td>
</tr>
<tr>
<td>Group number (≤4 vs&gt;4)</td>
<td>-.234*</td>
<td>.114</td>
<td>.047</td>
<td>-.553**</td>
</tr>
</tbody>
</table>

Note. Δ=difference; AFI=alternative fit indices, *p<0.05; **p<0.01
Figure 2.1 Number of studies on MI from 1991 to 2021
Note. In the 63 MI papers, 129 MI tests were conducted across different groups with MG-CFA.

Figure 2.2 Flow chart of the study selection process
CHAPTER 3: STUDY 2

FACTOR STRUCTURE AND CULTURAL MEASUREMENT INVARiance OF
THE STRENGTHS AND DIFFICULTIES QUESTIONNAIRES FOR
ELEMENTARY SCHOOL STUDENTS IN THE UNITED STATES AND CHINA

Abstract

The current study aimed to explore the factor structure and the measurement invariance of the teacher-reported Strengths and Difficulties Questionnaire (SDQ) across cultures (the United States and China). A sample of elementary school students in the U.S. (N=639) and China (N=666) was used. Confirmatory factor analysis results identified the underlying 3-factor structure of the SDQ for each sample (i.e., Internalizing Problems, Externalizing Problems, and Prosocial Behavior). A Multiple-group Confirmatory Factor Analysis indicated that SDQ showed full configural, metric, scalar, and residual invariance across the U.S. and China within the elementary school setting. Latent mean testing revealed that elementary school students in China were at a higher risk of externalizing problems, while U.S. elementary school children displayed a higher risk of prosocial behavior challenges. No significant differences in children’s internalizing problems between elementary school children in the U.S. and China. The present study’s findings contributed to the literature by adding additional validity evidence of teacher-reported SDQ.

Keywords: factor structure, latent means comparison, measurement invariance, validity
Introduction

The Strengths and Difficulties Questionnaire (SDQ) is a brief screening instrument designed to assess the behavioral attributes of children aged from 4 to 17 (Goodman, 2001). SDQ has an equivalent version for parents and teachers to assess children aged 3-16 years old and a near-identical version for youth aged 11-16-years-old to complete by themselves (Goodman, 1997; Goodman et al., 1998). All three versions of SDQ measure five dimensions of children’s behavioral attributes, including Emotional Problems, Conduct Problems, Hyperactivity/Inattention Problems, Peer Relationship Problems, and Prosocial Behavior. SDQ is available in more than 80 languages (SDQ, n.d.). The recommended criteria for selecting appropriate screening instruments for assessing children’s behavioral problems include minimal time investment, low-to-no cost, easy readability, multiple language choices, and feasibility (Furr, 2011; Harrison, 2013). Based on these criteria, SDQ was selected as the focus of this study. It has the following advantages: first, it is brief with only 25 items; second, it is freely available; third, the SDQ measures multiple domains of children’s behavioral problems. These advantages make the SDQ one of the most commonly used screening instruments worldwide. including Nordic countries (Obel et al., 2004; Smedje et al., 1999), Germany (Klasen et al., 2003), Netherlands (Van Widenfelt et al., 2003), and China (Du et al., 2008; Gao et al., 2013; Guo et al., 2018; Mellor et al., 2016).

As the SDQ is widely used, researchers investigated various psychometric properties of the SDQ, including factor structure (Bibou-Nakou et al., 2019; Kersten et al., 2016) and measurement invariance (MI) (Duinhof et al., 2020; Gomez & Stavropoulos, 2020; Janitza et al., 2020; Liang et al., 2019; Murray et al., 2021).
findings on the factor structure of the SDQ were mixed with the factor numbers ranging from 3 to 6. Further research was necessary to re-examine the factor structure of SDQ in different settings.

The previous research investigated the measurement invariance across different groups (e.g., informants, gender, age, race/ethnicity, income, level of education, parents, language versions, settings, and time) (Chiorri et al., 2016; DeVries et al., 2017; Gomez & Stavropoulos, 2020; He et al., 2013; Janitza et al., 2020; Liang et al., 2019; Murray et al., 2021; Rogge et al., 2018; Van de Looij-Jansen et al., 2011). To our knowledge, no study has examined the MI of the teacher version of the SDQ across China and the U.S. within elementary school contexts. The study on this group of students is significant as approximately 2.26 million American children aged 6-11 years had ongoing behavior problems in 2016 (Ghandour et al., 2019). In China, the mental issues of children and adolescents reached 17% (Dai et al., 2020). Social, emotional, and behavioral problems may negatively impact children’s social and academic competence (Midouhas et al., 2014; Romano et al., 2010; Waller et al., 2017), school engagement and behavioral engagement (Olivier et al., 2020), later education (Halle & Darling-Churchill, 2016), and even lead to more severe mental disorders in adulthood (Aebi et al., 2014; Cullins & Mian, 2015). Therefore, schools should focus on early identification and screening procedures to identify students at risk for mental health problems (Lane et al., 2012). Among diverse screening instruments, the SDQ has been investigated as a universal screener in China (Liang et al., 2019; Liu et al., 2013) and the U.S. (He et al., 2013; Palmieri & Smith, 2007). Besides, with the increasing educational exchange between China and U.S., more researchers seek to understand the differences in mental health
across these two cultures. In this case, it is incumbent on researchers to first verify that teachers from China and the U.S. interpret the SDQ items in the same way. This validity checking ensures that cross-cultural differences are not due to measurement variance. Therefore, the primary purpose of the present study was to explore the factor structure and MI of the teacher version of the SDQ with the samples of elementary school children from both China and the U.S. With that, the latent factor means were compared across both countries. The cross-country comparison can advance our understanding of elementary school children’s mental health and provide information for national and global intervention and prevention efforts.

**Literature Review**

**Factor Structure of the SDQ**

The initially proposed five-factor structure of the SDQ was supported using principal component analysis (Goodman, 1997; Goodman et al., 1998) as well as confirmatory factor analysis (CFA) (Di Riso et al., 2010; Essau et al., 2012; He et al., 2013; Kersten et al., 2016; Liang et al., 2019; Van Roy et al., 2008). The three-factor structure was also identified. It consisted of three dimensions: Internalizing Problems (comprising Emotional Problems and Peer Relationship Problems subscales), Externalizing Problems (comprising Hyperactivity/Inattention Problems and Conduct Problems subscales), and Prosocial Behavior (Bibou-Nakou et al., 2019; Di Riso et al., 2010; Goodman et al., 2010). The five-factor structure (i.e., Emotional Symptoms, Conduct Problems, Peer Problems, Hyperactivity, and Prosocial Behavior) was recommended for high-risk samples, while the three-factor structure was for community samples (Goodman & Goodman, 2009). Studies with Chinese samples identified a 5-
factor structure of self-reported SDQ (Liang et al., 2019; Liu et al., 2013; Yao et al., 2009) and a 4-factor structure with parent and teacher versions of SDQ (i.e., Internalizing Problems, Externalizing Problems, Inattention problems, and Prosocial behavior). (Liu et al., 2013). Studies of U.S. samples revealed a 3-factor structure (Dickey & Blumberg, 2004) and 5-factor structure of self-reported SDQ (He et al., 2013). Most of these previous studies sampled adolescents or children ranging from early childhood to adolescents.

The mixed findings indicated the necessity to examine the factor structure of SDQ with different samples. Most of the previous studies used self-reported or parent-reported SDQ. However, teachers are more likely to take a normative approach in judging a child’s behavior than the parents (Knold et al., 2004). Besides, elementary school students may not provide reliable ratings of their behaviors due to their limited cognitive abilities. Given the limitations of previous research, this study examined the internal structure of the teacher-reported SDQ with the samples of elementary school children from both China and the U.S.

**Measurement Invariance of the SDQ**

MI examines whether differences in observed variables result from the differences between groups rather than the differences in conceptualizations of the underlying construct (Finney & Davis, 2003). The analysis of MI of latent constructs is a precondition for latent mean comparison across groups or time (Davidov, 2011). Multiple-group confirmatory factor analysis (MG-CFA) is one of the most widely used methods to test MI due to its flexibility in examining every measurement parameter (i.e., factor structure, factor loadings, item intercept or threshold, and item residual variance) (Brown,
Four sequential invariance tests are involved in testing MI: configural invariance, metric invariance, scalar invariance, and measurement error invariance. Configural invariance signifies that participants from different groups based their interpretation of the items on the same theoretical groundings (Cheung & Rensvold, 2002; Schmitt & Kuljanin, 2008). The configural invariance model is the baseline model for the subsequent model testing. Metric invariance implies that the strength of the association between items and latent factors is the same across groups (Cheung & Rensvold, 2002; Schmitt & Kuljanin, 2008); the non-invariance suggests that the items may have different meanings across groups. The scalar invariance indicates that the extent to which participants endorse an item is equivalent across groups (Meredith, 1993; Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000). The measurement error invariance tests whether the variance in items unexplained by the factors differs across groups (Jung & Yoon, 2016). Each model is successively tested by posing equality restraints on model parameters. Among the four MI conditions, configural and metric invariance are pre-conditions for the equivalence of the structural relationship between the underlying constructs across groups. Scalar invariance is necessary for latent mean comparison across groups (Ployhart & Oswald, 2004).

With SDQ being translated into diverse versions and adapted to use in different cultures, it is of great necessity to maximize the equivalence of data from the different cultures. Cross-national differences in SDQ ratings may be due to cultural bias rather than comparable differences in mental health (Goodman et al., 2012). Therefore, establishing MI is essential for accurate inferences and interpretations of the data.
collected from different cultural contexts. However, few studies addressed whether the latent structure underlying the SDQ is invariant across countries. Measurement invariance testing of self-reported SDQ across European countries either did not establish measurement invariance (Essau et al., 2012) or found partial measurement invariance (Duinhof et al., 2020; Ortuño-Sierra et al., 2015). The study by Stevanovic et al. (2015) did not identify an acceptable model in countries from Europe, Asia, and Africa (Bulgaria, Croatia, India, Nigeria, Turkey, Indonesia, and Serbia). In addition, these studies on MI across cultures only explored the MI of the self-reported SDQ across European countries. There is a need to verify the previous findings on alternative cultural contexts (i.e., U.S. and China) as instruments developed based on Western theories may have a bias toward eastern culture. The meanings or contents of the SDQ may be interpreted differently due to unique experiences or perspectives specific to cultures. Besides, these previous studies sampled only adolescents. As behavioral problems are expected to decline as children age (Campbell, 2002), the SDQ should indicate different mental health problems in adolescents than in children. To address the research gap, the current study investigated the cultural MI of the teacher-reported SDQ across the U.S. and China within elementary school settings.

**Cross-cultural Comparison of Behavioral and Emotional Problems**

The cross-cultural comparisons of children and adolescents’ behavioral and emotional problems were made across seven countries (Verhulst et al., 2003) and twenty-four countries (Rescorla et al., 2007) with Youth Self-report, two countries with Teacher’s Report Form (Crijnen et al., 2000), and twelve cultures with parent-reported Child Behavior Checklist (Crijnen et al., 1999). One of the limitations of these studies
was that researchers made cross-cultural comparisons without testing the MI of the instruments they used. Based on our review, only two studies conducted the cross-cultural comparison of adolescents’ behavioral and emotional problems with Self-reported SDQ across European countries after testing MI (Duinhof et al., 2020; Ortuño-Sierra et al., 2015). Previous research did not compare the U.S. and Chinese elementary children’s behavioral and emotional problems assessed by teacher-reported SDQ. If the cultural measurement invariance of SDQ is established, the comparison will be made.

**Purposes and Research Questions**

Based on the previous studies, the current study was conducted with three aims. First, the previously proposed internal structure of teacher-reported SDQ in samples of elementary school students from the U.S. and China was examined. Second, the cultural MI of the teacher-reported SDQ across the U.S. and China was evaluated. Third, if the MI holds, differences in children’s mental health exist between the U.S. and China would be examined. As such, the current study sought to address the following questions:

1) What is the latent factor structure of teacher-reported SDQ for a sample of elementary school students from China and the U.S.?
2) Does cultural MI of teacher-reported SDQ hold across China and the U.S.?
3) Does elementary school children’s mental health differ across China and the U.S.?

**Method**

**Participants and Procedures**

From May to July of 2021, seventy-four public elementary school teachers from China and seventy-one public elementary teachers from the U.S. were invited to rate their students’ social, emotional, and behavioral problems (SEB) with SDQ by using the
Online Google Form. Each teacher selected 3 students potentially at low-risk level, 3 at a moderate level, and another 3 at higher risk levels of SEB based on their daily classroom observations of students’ SEB performance and then rated these 9 students with SDQ. Additional information was gathered about children’s gender (i.e., male and female), ethnicity (i.e., majority and minority), and special service referral (i.e., whether students were referred to any special services due to their SEB problems). The sample consisted of 1,305 children aged from 5 to 12 years old (M=9.2, SD=2.0). Teachers from the U.S.A rated 639 students, and teachers from China rated 666 students. Institutional Review Board approval informed consent was obtained before data collection. The ethical guidelines during the data collection and analysis procedures was followed.

Table 3.1 provides demographic information of the children rated in the sample. Both American and Chinese teachers rated slightly more male students. Most U.S. students came from minority groups, while most of the Chinese was from a majority group. Most of both samples had never received treatment due to their SEB problems. The mean age of both samples was approximately the same.

**Instrument**

SDQ is a 25-item questionnaire for assessing the SEB problems of children and adolescents aged from 4 to 17 years. It is a 3-point scale with anchors of “certainly true” =0, “somewhat true” =1, and “not true” =2. The scale intends to measure five dimensions of mental health problems: Emotional Problems (e.g., “I get a lot of headaches, stomach aches or sickness”), Conduct Problems (e.g., “I get very angry and often lose my temper”), Hyperactivity/Inattention Problems (e.g., “I am restless, I cannot stay still for long”), Peer Relationship Problems (e.g., “I would rather be alone than with people of my
age”) and Prosocial Behavior (e.g., “I am kind to younger children”). Subscale scale scores are computed by summing scores on relevant items after reverse coding positively worded items. Summing subscale scores yielded a total difficulty score ranging from 0 to 40. Higher difficulty scores reflect high-risk levels. The prosocial behavior subscale is not used for producing the total difficulties score. For teachers who completed SDQ, the sum scores fall into three categories based on an estimate of risk: normal (0-11), borderline (12-15), and abnormal (16-40). Higher scores on Prosocial Behavior reflect lower risk levels. The English version (Goodman, 1997) and Chinese versions of the SDQ teacher-report (Du et al., 2008) were used for the U.S. sample and Chinese sample, respectively.

**Data Analysis**

All statistical analyses were conducted using Mplus 8.4 (Muthén & Muthén, 2019). The weighted least squares with mean and variance adjusted (WLSMV) estimation method was used to accommodate the categorical nature of the data (Finney & DiStefano, 2013). The theta parameterization was used for all models to allow residual variance to be parameters (Gunzler et al., 2021). No missing data were identified. The clustering design effects resulting from the nesting of students within teachers were considered to provide more accurate parameter estimates as recommended by Raykov and DiStefano (2021). Cronbach’s alpha reliability coefficients of subscales were calculated in the U.S. sample, Chinese samples, and overall sample.

Confirmatory factor analysis (CFA) was conducted to examine whether teacher ratings of elementary school children from the U.S. and China would support the previously identified factor structure. The 3-factor structure (i.e., Internalizing Problems,
Externalizing Problems, and Prosocial Behavior) for the community sample, the 5-factor structure (i.e., Emotional Problems, Conduct Problems, Hyperactivity/Inattention Problems, Peer Relationship Problems, and Prosocial Behavior) for high-risk samples (Goodman & Goodman, 2009) and 4-factor structure (Internalizing Problems, Externalizing Problems, Inattention Problems, and Prosocial Behavior) (Liu et al., 2013). To make the results interpretable, the items on Prosocial Behavior were worded in the same direction as those on the other four subscales so that higher scores indicated more prosocial problems. The factor structures for the Chinese sample and the American sample were examined. A common measurement model is established only if the proposed model fits both samples.

The model fit was examined with the following indices commonly used with the analysis of categorical data (Finney & DiStefano, 2013): the $p$-value associated with the WLSMV-based Chi-square statistic fit statistic, comparative fit index (CFI), root mean squared error of approximation (RMSEA), and Standardized Root Mean Square (SRMR). As chi-square is sensitive to large sample sizes or complex models, additional fit indices are often used in conjunction with a significant value to provide additional information about the degree of model fit or misfit (Brown, 2015). The following cutoff values suggested an acceptable model fit: CFI $\geq .90$, RMSEA $\leq .08$, and SRMR $\leq .10$. The higher cutoff values indicated good model fit: CFI $\geq .95$, RMSEA $\leq .05$, and SRMR $\leq .08$ (Hu & Bentler, 1999). As global model fit might be affected by the magnitude of the factor loadings and the number of items (Greiff & Heene, 2017; McNeish et al., 2018), the local fit indices were also examined. Costello and Osborne (2005) suggested that factor loadings of the items should be greater than 0.30. Standardized residuals greater
than |3.0| indicate a possible local model misfit (Raykov & Marcoulides, 2012). If models did not show acceptable model fits, model modification indices were consulted to identify the misspecified parameters.

The configural, metric, scalar invariance, and residual invariance were sequentially tested using an MG-CFA model. Configural invariance was analyzed by constraining the factor structure to be identical across the U.S. and China. In this model, the factor variance was fixed to 1, and the factor mean was fixed to 0 in each group for model identification so that all factor loadings and thresholds were then estimated. The residual variance was constrained to 1 in both groups as the residual variances are not uniquely identified in the configural invariance model. Once the configural invariance was established, the metric invariance was tested by constraining the unstandardized factor loadings to be equal across both countries. In testing metric invariance, the factor variance was fixed to 1 in the U.S. group for identification but was freely estimated in the Chinese group. The factor mean was fixed to 0 in both groups for identification. All unstandardized factor loadings were constrained to be equal across the two countries. All item thresholds were estimated, and all residual variances were constrained to 1 across two countries. After establishing the metric invariance, the scalar invariance was tested in this model, the latent factor mean and variance were fixed to 0 and 1, respectively, in the U.S. group, but they were freely estimated for the Chinese group. All factor loadings and item thresholds were constrained to be equal across the two countries. All residual variances were constrained to be 1 in both groups. To test residual variance invariance, two models were compared. First, a model (i.e., residual variance invariance model A) was fit in which all the residual variances were freely estimated in the Chinese group.
The residual variance invariance model A was then compared with residual variance invariance model B, in which all residual variances were fixed to 1 in the Chinese group. The residual variances in the U.S. group were all fixed to 1 for model identification in both models. The remaining model parameters were estimated as with the scalar invariance model. If invariance cannot be established at one of these successive steps, partial invariance can be tested by freeing factor loading or intercept/threshold of one item at a time by consulting modification indices (Dimitrov, 2010).

As WLSMV-based chi-square statistics across models are of different scales, the DIFFTEST option in Mplus was used to evaluate the chi-square change ($\Delta \chi^2$) for statistical significance relative to the change in the degree of freedom for the nested model comparison. Although the $\Delta \chi^2$ test allows the statistical nest model comparison, it has some drawbacks. First, $\Delta \chi^2$ is sensitive to departures from multivariate normality. Second, $\Delta \chi^2$ is always large and statistically significant with complex models and/or large samples (Chen, 2007). In addition to statistical model fit indices ($\Delta \chi^2$), the difference between models’ RMSEA values ($\Delta$ RMSEA), CFI values ($\Delta$ CFI), and SRMR values ($\Delta$ SRMR) was used to assess the measurement invariance. When the changes in these fit indices for the more restricted model met the criteria ($\Delta$ CFI$\leq$0.01, $\Delta$ RMSEA$\leq$0.015, and $\Delta$ SRMR$\leq$0.03 for the factor loadings invariance, and $\Delta$ CFI$\leq$0.01, $\Delta$ RMSEA$\leq$0.015, and $\Delta$ SRMR$\leq$0.01 for the item intercepts/thresholds and residual invariance given sample size $>300$), the hypothesis of invariance was tenable (Chen, 2007). In the case of a non-invariance, modification indices were consulted to determine which factor loading(s) or threshold(s) should be freely estimated for partial invariance.
If the scalar invariance were established, the latent means across the U.S. and Chinese samples would be examined. The means of the latent variables in the U.S. group were fixed to 0, and those in the Chinese group were freely estimated. The size of the standardized latent mean differences was estimated using Cohen’s $d$.

**Results**

**Confirmatory Factor Analysis**

The CFA analysis results for the U.S. and Chinese samples are presented in Table 3.2. The five-factor solution was first examined. The model for the U.S. sample showed an adequate fit. However, the model for the Chinese sample did not work well due to the lack of convergence. four-factor solution was then examined. The solution did not fit the data for the U.S. or Chinese sample with all the model fit indices outside the recommended boundaries. The three-factor structure of the SDQ (Internalizing Problems, Externalizing Problems, and Prosocial Behavior) was finally examined. Model 1 for the U.S. sample did not fit well; the RMSEA model fit index was outside the recommended boundary. The CFI value was outside the recommended boundaries for the Chinese sample. The MI suggested that the model misfit involved Item 11 and Item 14 with loadings on different factors. In Model 2, Item 11 (“I have one good friend or more.”) and Item 14 (“Other people my age generally like me.”) was allowed to load on Prosocial Behavior rather than on Internalizing Problems. The model fit improved with all the indices within the recommended boundaries for all two samples. The local model fit information was then checked, and results showed that the loading values of Item 23 (“I get along better with adults than with people my own age.”) were lower than 0.30 for all three samples (i.e., 0.22, -0.05, and 0.12 respectively for the U.S., Chinese, and overall
sample), which indicated that this item failed to measure the constructs adequately. Therefore, in model 3, Item 23 was removed from model 2. Even if the global model fit was almost the same as that for Model 2, the local model fit is good with all the factor loadings greater than 0.50 for each sample. Figure 3.1 presents the factor structure of the final optimal model.

As presented in Table 3.3, standardized factor loadings indicated significant associations of all items with the stated factors. All the loading values are positive, indicating a positive linear relationship between the items and factors. The correlation between the three factors measured by teacher-reported SDQ varied from 0.31 to 0.77, as shown in Table 3.4. Internalizing Problems were highly and positively correlated with Externalizing Problems for three samples, with correlation coefficients ranging from 0.47 to 0.73, indicating that children at a higher risk of internalizing problems are more likely to have externalizing problems. The correlation is stronger for the Chinese sample than for the U.S. sample. The correlation coefficient between Prosocial Behavior and Internalizing Problems or Externalizing Problems ranges from 0.31 to 0.77, indicating that children with more positive prosocial behaviors tend to be at lower risk levels for internalizing and externalizing problems. The correlation in the U.S. sample is stronger

**Reliability**

Cronbach’s alpha reliability coefficients for all factors in the Chinese sample (Internalizing Problems: \( \alpha = 0.89 \), Externalizing Problems: \( \alpha = 0.86 \), Prosocial Behavior: \( \alpha = 0.89 \)) and the U.S. sample (Internalizing Problems: \( \alpha = 0.82 \), Externalizing Problems: \( \alpha = 0.90 \), Prosocial Behavior: \( \alpha = 0.89 \)) were greater than 0.80. Cronbach’s alpha reliability
coefficients for all factors in the overall sample were greater than 0.80 (Internalizing Problems: $\alpha=0.86$, Externalizing Problems: $\alpha=0.88$, Prosocial Behavior: $\alpha=0.89$).

**Measurement Invariance Testing**

Table 3.5 summarizes the results of the series of MI tests with MG-CFA. Configural invariance is the three-factor structure identified in the previous CFA analysis, and it was used across the U.S. and Chinese samples. It showed an acceptable model fit ($\text{CFI}=0.924$, $\text{RMSEA}=0.069$, $\text{SRMR}=0.126$), indicating that the same internal structure of teacher-reported SDQ was identified across the U.S. and China. The metric invariance model was compared with the configural model. Although the scaled chi-square difference test was statistically significant, the changes in the alternative measures of fit indicated that the hypothesis of equal loadings was tenable ($\Delta\text{CFI}=0.001$, $\Delta\text{RMSEA}=0.001$, $\Delta\text{SRMR}=0.007$). Next, the scalar invariance was tested by comparing this model with the metric invariance model. The change in chi-square was statistically significant. However, the alternative measures of fit changes indicated that the hypothesis of equal thresholds for all the items was tenable ($\Delta\text{CFI}=0.005$, $\Delta\text{RMSEA}=0.001$, $\Delta\text{SRMR}=0.001$). Although the chi-square difference between the residual variance invariance model A and model B was significant, the alternative measures of fit indicated that equal residual variances for the items were tenable ($\Delta\text{CFI}=0.002$, $\Delta\text{RMSEA}=0.001$, $\Delta\text{SRMR}=0.007$), supporting the hypothesis of equal error variances for the items.

As residual invariance of the teacher-reported SDQ was supported, latent factor means across the U.S. and China were compared. The estimated factor means for Externalizing Problems and Prosocial Behavior in the Chinese group are statistically different from 0, indicating that the latent factor means for these two subscales are
unequal across the U.S. and Chinese samples. Students in China were rated to be at a higher risk level of Externalizing problems ($d=0.21$, SE= 0.09, $p<0.05$) but lower risk level of Prosocial Behavior relative to students in the U.S. ($d=-0.23$, SE= 0.11, $p<0.05$). The estimated factor means for Internalizing Problems were not statistically different between the U.S. and Chinese students.

**Discussion**

While SDQ is commonly used in research, clinical, and community practice; the ongoing evaluations of the psychometric properties of the SDQ are necessary as an instrument’s effectiveness depends mainly on its reliability and validity. Although prior studies were conducted to investigate the psychometric quality of the SDQ, the findings of the internal structure of the SDQ are inconsistent in different settings (e.g., Bibou-Nakou et al., 2019; Kersten et al., 2016; Liang et al., 2019), and the practice of evaluating cross-cultural MI of the SDQ is limited. Besides, most previous studies focused on adolescents instead of young children, while the current study sampled teacher ratings of elementary school children. As health problems may negatively impact parental relationships, physical health, and participation in society if left unaddressed (Goodman et al., 2011; Pagliaccio et al., 2012; Shonkoff et al., 2012), school-based universal screening of children for mental health problems was recommended for identifying children at risk at an early stage (Dever et al., 2015; Glover & Albers, 2007; Kettler et al., 2014). In this context, choosing a valid screening instrument is especially important. Therefore, we examined the factor structure and measurement invariance of the teacher-reported SDQ to add to the validity evidence of the SDQ.
The CFA results supported the scale’s underlying 3-factor structure identified in the previous studies (Bibou-Nakou et al., 2019; Dickey & Blumberg, 2004; Di Riso et al., 2010; Goodman et al., 2010; Koskelainen et al., 2001). The CFA model with three latent factors (i.e., Internalizing Problems, Externalizing Problems, and Prosocial Behavior) yielded an acceptable model fit in the overall sample and those from the U. S. and China. Item 23 (“I get along better with adults than with people my own age.”) failed to explain the latent construct of Internalizing problems. The failure might result from teachers having insufficient access to compare how well an individual child interacts and communicates with adults as compared with other children of his/her own age. Removing this item while screening elementary school students with teacher-reported SDQ in the U.S. or China was suggested. Besides, Item 11 (“I have one good friend or more.”) and Item 14 (“Other people my age generally like me.”) loaded on Prosocial Behavior rather than Internalizing Problems. The finding corresponds to the theory that children with stronger social skills are more likely to engage in prosocial behaviors (Sutton et al., 1999). The same results were also identified by Di Riso et al. (2010). The consistency indicates that these two items explain the construct of Prosocial Behavior better. These two items were recommended to assess children’s prosocial behavior rather than internalizing problems within elementary school settings in both the U.S. and China.

For researchers interested in a cross-cultural comparison of the mental health measured by teacher-reported SDQ between the two countries, MI test was conducted to examine the differences between the U.S. and Chinese samples in the measurement properties, which included factor loadings, item threshold, and residual variances. The results of the MG-CFA supported the configural, metric, scalar invariance, and residual
variance invariance for the teacher-reported SDQ (i.e., equality of factor patterns, factor loadings, and the equality of thresholds and residual variances). The full invariance results indicated that children’s mental health measured by teacher-reported SDQ could be compared across the U.S. and China within elementary school settings. To our knowledge, this study is the first to investigate the measurement invariance of the SDQ across the U.S. and China. Therefore, this study contributed to the research on the psychometric properties of the SDQ by providing additional evidence of the SDQ’s validity.

The latent means comparison across the U.S. and China showed that students in China had higher risk levels of externalizing problems. A previous study showed that lower risk levels of externalizing problems were reported in non-western cultures (Lam et al., 2001). The inconsistency might be due to the sampling differences. The current findings implied that Chinese elementary schools should pay more attention to children’s externalizing problems, as externalizing problems have a negative impact on academic performance, parenting, and delinquency (Deighton et al., 2018; Fergusson et al., 2005; Scaramella et al., 2008; Serbin et al., 2015; Vassallo et al., 2002).

The U.S. elementary school students were rated to have a higher risk level of prosocial behavior than students from China, indicating that U.S. elementary schools should foster children’s prosocial tendencies. Prosocial behaviors help ease children’s stress and increase their sense of meaning and purpose and self-efficacy (Raposa et al., 2016). No significant differences in children’s internalizing problems between elementary school children in the U.S. and China were found. This finding differed from the previous study, which indicated the Chinese sample had higher levels of depressive
symptoms than the U.S. sample (Stewart et al., 2004). The inconsistency might be due to the construct difference between depression and internalizing behavior. Our findings added to the present literature on the cross-cultural comparison of children’s mental health.

Despite its values and contributions, the current study has several limitations. First, the current study collected data only from a sample of children aged from 5 to 12 rated by teachers. The results may not apply to children of older age groups or in clinical settings. Future studies may investigate the measurement invariance of SDQ in other settings to examine whether the same results apply. Second, a three-factor structure was identified with one item removed due to the low factor loadings. Replication is needed to determine whether similar factor loadings can be identified in the sample from other countries, as well as samples from other age groups. Third, the current research examined the MI of the teacher-reported SDQ across the U.S. and China. As teacher raters vary in gender, educational background, years of teaching experience, and subject teaching, these factors may also impact their ratings. Future research may explore the MI across teachers concerning their characteristics within a culture. Fourth, the current study identified cross-cultural differences in externalizing problems and prosocial behaviors without exploring the cause of the differences. Future research may investigate the factors associated with such differences (e.g., gender, age, ethnicity, SES, and cultural variables).

Findings from the current study have implications for research and practice. The present study identified the three-factor internal structure of the teacher-reported SDQ: Internalizing Problems, Externalizing Problems, and Prosocial Behavior, providing evidence for the construct validity of the SDQ when used to screen elementary school
students’ mental health in the U.S. and China. The establishment of the MI across the U.S. and China implied that researchers could compare elementary school children’s mental health rated with the teacher-reported SDQ across the U.S. and China as teachers from the two countries perceived the meaning of mental health problems measured by SDQ in the same way. The findings of the present study contributed to the literature by adding additional validity evidence of teacher-reported SDQ, providing support for future researchers who are interested in cross-cultural comparisons of children’s mental health measured by teacher-reported SDQ. U.S. elementary school students have higher risk levels of prosocial behavior but lower levels of externalizing problems than the Chinese. The findings provided insight into both U.S. and Chinese elementary schools for their intervention and prevention efforts.

References


[https://doi.org/10.1207/s15328007sem0902_5](https://doi.org/10.1207/s15328007sem0902_5)

[https://doi.org/10.1177/1073191114568301](https://doi.org/10.1177/1073191114568301)


[https://doi.org/10.1176/ajp.156.4.569](https://doi.org/10.1176/ajp.156.4.569)

[https://doi.org/10.1034/j.1600-0447.2000.102006439.x](https://doi.org/10.1034/j.1600-0447.2000.102006439.x)


[https://doi.org/10.1093/ijpor/edq031](https://doi.org/10.1093/ijpor/edq031)

[https://doi.org/10.1111/bjdp.12218](https://doi.org/10.1111/bjdp.12218)

[https://doi.org/10.1002/pits.21825](https://doi.org/10.1002/pits.21825)


Rescorla, L., Achenbach, T. M., Ivanova, M. Y., Dumenci, L., Almqvist, F., Bilenberg, N., ... & Verhulst, F. (2007). Epidemiological comparisons of problems and positive qualities reported by adolescents in 24 countries. *Journal of consulting and clinical psychology, 75*(2), 351. [https://doi.org/10.1037/0022-006x.75.2.351](https://doi.org/10.1037/0022-006x.75.2.351)


effects and minor factors?. *British Journal of Clinical Psychology, 50*(2), 127-144. [https://doi.org/10.1348/014466510x498174](https://doi.org/10.1348/014466510x498174)


Table 3.1 Demographic characteristics of participants

<table>
<thead>
<tr>
<th>Variables</th>
<th>U.S. Sample (N=639)</th>
<th>China Sample (N=666)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>355 (55.6%)</td>
<td>386 (58.0%)</td>
</tr>
<tr>
<td>Female</td>
<td>283 (44.4%)</td>
<td>280 (42.0%)</td>
</tr>
<tr>
<td>Ethnicity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Majority</td>
<td>276 (43.2%)</td>
<td>659 (98.9%)</td>
</tr>
<tr>
<td>Minority</td>
<td>363 (56.8%)</td>
<td>7 (1.1%)</td>
</tr>
<tr>
<td>Special Service</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>475 (74.3%)</td>
<td>588 (88.3%)</td>
</tr>
<tr>
<td>Yes</td>
<td>164 (25.7%)</td>
<td>78 (11.7%)</td>
</tr>
<tr>
<td>Mean (SD)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>8.9 (2.0)</td>
<td>9.5 (1.9)</td>
</tr>
</tbody>
</table>
### Table 3.2 CFA model fit statistics

<table>
<thead>
<tr>
<th>Model</th>
<th>5 factors</th>
<th>4 factors</th>
<th>3 factors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>U.S. sample</td>
<td>China sample</td>
<td>U.S. sample</td>
</tr>
<tr>
<td></td>
<td>0.949</td>
<td>0.697</td>
<td>0.803</td>
</tr>
<tr>
<td></td>
<td>0.943</td>
<td>0.664</td>
<td>0.781</td>
</tr>
<tr>
<td></td>
<td>0.070(0.065-0.074)</td>
<td>0.168(0.164-0.172)</td>
<td>0.086(0.082-0.090)</td>
</tr>
<tr>
<td></td>
<td>0.098</td>
<td>0.190</td>
<td>0.143</td>
</tr>
</tbody>
</table>

**Note.** $\chi^2$=Chi-square test statistic; df= degree of freedom; CFI=comparative fit index; TLI= Tucker-Lewis Index; RMSEA= root mean squared error of approximation; CI= confidence interval; SRMR= standardized root mean square residual.
Table 3.3 Three-factor CFA results: standardized factor loadings for group samples

<table>
<thead>
<tr>
<th>Factors and Items</th>
<th>U.S. sample</th>
<th>Chinese sample</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Internalizing Problems</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. I get a lot of headaches, stomach-aches or sickness.</td>
<td>0.72</td>
<td>0.78</td>
</tr>
<tr>
<td>6. I would rather be alone than with people of my age.</td>
<td>0.60</td>
<td>0.81</td>
</tr>
<tr>
<td>8. I worry a lot.</td>
<td>0.70</td>
<td>0.85</td>
</tr>
<tr>
<td>13. I am often unhappy, depressed or tearful.</td>
<td>1.00</td>
<td>0.89</td>
</tr>
<tr>
<td>16. I am nervous in new situations. I easily lose confidence.</td>
<td>0.70</td>
<td>0.84</td>
</tr>
<tr>
<td>19. Other children or young people pick on me or bully me.</td>
<td>0.76</td>
<td>0.82</td>
</tr>
<tr>
<td>24. I have many fears, I am easily scared.</td>
<td>0.68</td>
<td>0.76</td>
</tr>
<tr>
<td><strong>Externalizing Problems</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. I am restless, I cannot stay still for long.</td>
<td>0.80</td>
<td>0.82</td>
</tr>
<tr>
<td>5. I get very angry and often lose my temper.</td>
<td>0.84</td>
<td>0.81</td>
</tr>
<tr>
<td>10. I am constantly fidgeting or squirming.</td>
<td>0.80</td>
<td>0.82</td>
</tr>
<tr>
<td>12. I fight a lot. I can make other people do what I want.</td>
<td>0.85</td>
<td>0.85</td>
</tr>
<tr>
<td>15. I am easily distracted. I find it difficult to concentrate.</td>
<td>0.85</td>
<td>0.85</td>
</tr>
<tr>
<td>18. I am often accused of lying or cheating.</td>
<td>0.84</td>
<td>0.83</td>
</tr>
<tr>
<td>22. I take things that are not mine from home, school or elsewhere.</td>
<td>0.74</td>
<td>0.71</td>
</tr>
<tr>
<td>25. I finish the work I'm doing. My attention is good.</td>
<td>0.84</td>
<td>0.72</td>
</tr>
<tr>
<td><strong>Prosocial Behavior</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. I try to be nice to other people. I care about their feelings.</td>
<td>0.87</td>
<td>0.84</td>
</tr>
<tr>
<td>4. I usually share with others, for example CD’s, games, food.</td>
<td>0.80</td>
<td>0.78</td>
</tr>
<tr>
<td>7. I usually do as I am told.</td>
<td>0.89</td>
<td>0.62</td>
</tr>
<tr>
<td>9. I am helpful if someone is hurt, upset or feeling ill.</td>
<td>0.83</td>
<td>0.78</td>
</tr>
<tr>
<td>11. I have one good friend or more.</td>
<td>0.74</td>
<td>0.75</td>
</tr>
<tr>
<td>14. Other people my age generally like me.</td>
<td>0.85</td>
<td>0.87</td>
</tr>
<tr>
<td>17. I am kind to younger children.</td>
<td>0.76</td>
<td>0.83</td>
</tr>
<tr>
<td>20. I often offer to help others (parents, teachers, children).</td>
<td>0.81</td>
<td>0.82</td>
</tr>
<tr>
<td>21. I think before I do things.</td>
<td>0.83</td>
<td>0.77</td>
</tr>
</tbody>
</table>
### Table 3.4 Correlations among teacher-reported SDQ factors

<table>
<thead>
<tr>
<th>Sample</th>
<th>Externalizing Problems &amp; Internalizing Problems</th>
<th>Prosocial Behavior &amp; Externalizing Problems</th>
<th>Prosocial Behavior &amp; Internalizing Problems</th>
</tr>
</thead>
<tbody>
<tr>
<td>U.S. Sample</td>
<td>0.41</td>
<td>0.54</td>
<td>0.37</td>
</tr>
<tr>
<td>China Sample</td>
<td>0.50</td>
<td>0.39</td>
<td>0.18</td>
</tr>
</tbody>
</table>

### Table 3.5 MG-CFA model fit across the U.S. and Chinese samples

<table>
<thead>
<tr>
<th>Model</th>
<th>$\chi^2$(df)</th>
<th>$\Delta$SB- $\chi^2$(Δ df)</th>
<th>CFI</th>
<th>$\Delta$CFI</th>
<th>RMSEA[90%CI]</th>
<th>$\Delta$RMSEA</th>
<th>SRMR</th>
<th>$\Delta$SRMR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Configural invariance</td>
<td>1763.845*(490)</td>
<td>____</td>
<td>0.937</td>
<td>____</td>
<td>0.063[0.060-0.066]</td>
<td>____</td>
<td>0.095</td>
<td>____</td>
</tr>
<tr>
<td>Metric invariance</td>
<td>1717.987*(511)</td>
<td>88.801*(21)</td>
<td>0.941</td>
<td>0.004</td>
<td>0.060[0.057-0.063]</td>
<td>-0.003</td>
<td>0.101</td>
<td>0.006</td>
</tr>
<tr>
<td>Scalar Invariance</td>
<td>1808.553(559)</td>
<td>156.370*(48)</td>
<td>0.938</td>
<td>-0.003</td>
<td>0.059[0.056-0.062]</td>
<td>-0.001</td>
<td>0.104</td>
<td>0.003</td>
</tr>
<tr>
<td>Residual Variance Invariance</td>
<td>1876.737*(532)</td>
<td>____</td>
<td>0.934</td>
<td>____</td>
<td>0.062[0.059-0.065]</td>
<td>____</td>
<td>0.096</td>
<td>____</td>
</tr>
<tr>
<td>Residual variance invariance A</td>
<td>1870.869*(556)</td>
<td>120.224*(24)</td>
<td>0.935</td>
<td>0.001</td>
<td>0.060[0.057-0.063]</td>
<td>-0.002</td>
<td>0.103</td>
<td>0.006</td>
</tr>
</tbody>
</table>

Note. MG-CFA=multigroup confirmatory factor analysis; $\chi^2$=Chi-square test statistic; df= degree of freedom; SB-$\chi^2$=Satorra-Bentler scaled chi-square; CFI=comparative fit index; RMSEA= root mean square error of approximation; CI= confidence interval; SRMR= standardized root mean square residual; Δ=difference.
*p<.001.
Note. prosocial=Prosocial Behavior; externalizing=Externalizing Problems; Internalizing=Internalizing Problems.

Figure 3.1 The three-factor structure of teacher-reported SDQ
CHAPTER 4: STUDY 3
THE IMPACT OF GROUP SIZE RATIO AND MODEL SIZE ON THE
SENSITIVITY OF FIT MEASURES IN MEASUREMENT INVARIANCE
TESTING: A MONTE CARLO SIMULATION STUDY

Abstract
A Monte Carlo simulation study was conducted to evaluate the impact of group size ratio in combination with model size on the sensitivity of the three fit measures (i.e., ΔCFI, ΔRMSEA, and ΔSRMR), which are commonly used to detect a lack of metric invariance or scalar invariance. Design factors included model size, group size ratio, location of noninvariance, and location of noninvariant items. Results suggested that CFI failed to detect scalar noninvariance when invariance tests involved a group ratio of 1:4 in combination with the location of non-invariant items only on one factor. RMSEA failed to detect either metric or scalar invariance with a large model size (e.g., 2 factors/ 30 indicators), especially when the group size ratio was 1:4 and the location of non-invariant items on one factor was involved. SRMR failed to detect metric invariance when a group ratio of 1:2 was involved. SRMR was not adequate for detecting scalar invariance.
Recommendations for empirical researchers are provided based on the findings.

Keywords: fit measures, measurement invariance, model size, group size ratio, sensitivit
Introduction

Measurement invariance (MI) refers to how instruments administered under in different conditions yield identical psychometric properties (Cheung & Rensvold, 1999). The procedures examine whether differences in observed variables are a function of the differences between groups rather than the differences in conceptualizations of the underlying construct (Finney & Davis, 2003). MI is an essential psychometric property, and the lack of MI implies that the same instrument does not measure the same underlying constructs across conditions (Cieciuch et al., 2019; Millsap & Kwok, 2004). Researchers who conduct comparative studies across groups or measurement occasions should ensure that the differences in the latent construct arise from the differences in the observed scores. Thus, establishing MI is a pre-condition for group comparisons (Borsboom, 2006; Davidov, 2011; Meredith & Teresi, 2006).

Researchers have realized the importance of MI tests for making valid conclusions about the group differences in latent constructs. In a recent search of the Education Resources Information Center (ERIC) database with delimiters “Measurement invariance”, the number of studies on MI increased from 1,541 in 2015 to 4,506 in 2021. This demonstrates a dramatic increase of MI investigations on the field of education.

Multiple-group confirmatory factor analysis (MGCFA) is a common method for testing MI due to its flexibility in examining individual parameters for a better understanding of similarities and differences across different conditions using the same instrument (Brown, 2015; Jung & Yoon, 2016; Schmitt & Kuljanin, 2008; Vandenberg & Lance, 2000). To determine MI, researchers rely on different model fit statistics provided by the CFA models across groups. These indices typically include the overall Chi-square
test of exact fit, Root Mean Square Error of Approximation (RMSEA), Comparative Fit Index (CFI), Tucker-Lewis Index (TLI, or the Non-Normed Fit Index), and Standardized Root Mean Square Residual (SRMR) (DiStefano & Hess, 2005; Jackson et al., 2009; Kline, 2015). After testing configural invariance, additional tests of invariance were examined by comparing the model fit indices between two nested models. The change in chi-square (Δχ²), CFI (ΔCFI), RMSEA (ΔRMSEA), and SRMR (ΔSRMR) have been recommended to examine invariance as MI needs to be evaluated in a nested modeling framework. Nested model testing examines whether differences exist between models with more parameter constraints imposed and models with fewer constraints on the model parameters (e.g., configural vs. metric model) (Chen, 2007; Cheung & Rensvold, 2002; Meade et al., 2008; Rutkowski & Svetina, 2017; Vandenberg & Lane, 2000).

Various cutoff values for fit statistics have been suggested to evaluate MI. The review of literature results presented in Study 1 showed that 79.8% of the MI tests used ΔCFI, 67.1% used ΔRMSEA, and 27.1% used ΔSRMR. Different ΔCFI cutoff values were used in the literature, including 0.01 (83.5%), 0.02 (2.9%), 0.005 (5.8%), and 0.01 for metric invariance and 0.02 for scalar invariance (7.8%). Regarding ΔRMSEA cutoff values, 72.4% of the MI tests used a value of 0.015, 17.2% used 0.01, 9.2% used 0.015 for metric invariance and 0.03 for scalar invariance, and 1.1% used 0.015 for metric invariance and 0.03 for scalar invariance. Concerning ΔSRMR cutoff values, 51.4% of the MI used the values of 0.03, 11.4% used 0.015, 22.9% used 0.01 for metric invariance and 0.03 for scalar invariance, and 14.3% used 0.015 for metric invariance and 0.03 for scalar invariance. Overall, most studies used ΔCFI ≤ 0.010, ΔRMSEA ≤ 0.015, and ΔSRMR ≤ 0.30 as the criteria for evaluating MI.
Applied researchers used the criteria for fit measures mentioned above to evaluate MI regardless of the group and/or model sizes. However, different model sizes were involved in empirical research on MI tests. Study 1 showed that across 129 group comparisons, sample size ratios ranged from 1:1 to 1:39, including sample size ratios of 1:1 (67.9%), 1:2 (18.9%), 1:3 (2.8%), 1:4 (0.9%), 1:5 (4.7%), 1:6 (1.9%) and other ratios (2.7%) including 1:8, 1:28, and 1:39. Also, previous research has showed that group size ratio affects the performance of fit measures in detecting MI. For example, large imbalances in group size affect the MI tests results and led to inaccurate conclusions of invariance as groups with large sample size have more weight in deciding MI relative to small group sample size (Yoon & Lai, 2018). The power of fit measures for detecting noninvariance was found to decrease when the sample size differences between the two groups were large (Chen, 2007).

In addition to different group size ratios, different model sizes were involved in MI tests. Study 1 showed that the range of factors involved in MI tests was from 1 to 10 and that for items was from 3 to 87. The impact of model sizes on the performance of fit measures was also identified in previous research. For example, in the presence of model misspecification, RMSEA values decreased when the number of manifest variables increased while the performance of CFI was relatively stable (Breivik & Olsson, 2001; Kenny & McCoach, 2003).

Previous research explored the impact of model size (e.g., Cao & Liang, 2022) or group size imbalance (e.g., Yoon & Lai, 2018) on the sensitivity to fit measures in detecting MI. However, there is a gap in the literature concerning the effect of group size ratios in combination with model sizes on the sensitivity of fit measures in evaluating MI.
Therefore, the current study aimed to examine the impact of the model sizes and the group size ratios on the sensitivity of the model fit statistics ($\Delta$CFI, $\Delta$RMSEA, and $\Delta$SRMR) for detecting the metric and scalar MI. Specifically, the current study was guided by the following research questions:

1. In multiple-group confirmatory factor analysis model, what is the ability of $\Delta$CFI to detect metric or scalar noninvariance?

2. In multiple-group confirmatory factor analysis model, what is the ability of $\Delta$RMSEA to detect metric or scalar noninvariance?

3. In multiple-group confirmatory factor analysis model, what is the ability of $\Delta$SRMR to detect metric or scalar noninvariance?

**Literature Review**

**Measurement Invariance Tests with Multiple-Group CFA**

MI examines whether scores from the operationalization of a construct have the same meaning across different groups or conditions (Meade & Lautenschlager, 2004). Mathematically, MI refers to the conditional probability of having an observed score if a person’s ability is independent of group membership (Meredith & Millsap, 1992; Yoon & Millsap, 2007). MI is shown in the following equation: $P(X|W, G) = P(X|W)$, where $X$ is the observed score obtained from a measure, $W$ is the latent construct that is measured by the instrument, and $G$ refers to the group membership (Mellenbergh, 1989). The equation denotes that persons with identical abilities on a measured construct have the same probability of obtaining the observed scores regardless of a group membership. The establishment of the MI implies that a construct has the same meaning to different groups or across repeated measurement time points (Putnick & Bornstein, 2016). Ignoring MI
makes it impossible to make valid comparisons of the latent factor mean difference across groups, as observed differences across groups might not reflect the true differences in latent factors if a certain type of MI does not hold (Raju et al., 2002; Van De Schoot et al., 2015).

Four steps are involved in MI tests with MG-CFA approach. First, the configural invariance is tested by loading the same items on the same factor(s) across each group. The next step is metric invariance which evaluates invariant factor loadings between groups. The third step is scalar invariance which examines whether item intercepts/ thresholds were invariant in addition to factor loadings. The last step is residual invariance which evaluates whether the item residual variance is the same across groups in addition to the factor loadings and intercepts (Meredith, 1993; Vandenberg & Lance, 2000). As scalar invariance is needed for latent mean comparison across groups (Millsap, 2011; Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000), the current study focuses only on the power for detecting metric and scalar MI under different conditions.

The Impact of Group Size Ratio on Fit Measures in Measurement Invariance Testing

Previous research has investigated the changes in model fit statistics (e.g. $Δχ^2$, $Δ$CFI, $Δ$RMSEA, and $Δ$SRMR) for detecting MI (Chen, 2007; Cheung & Rensvold, 2002; French & Finch, 2006; Meade et al., 2008; Rutkowski & Svetina, 2014; Rutkowski & Svetian, 2017). Cutoff criteria for fit measures were inconsistent due to different factors. One of the factors examined by previous research was the group size ratio (Chen, 2007; Yoon & Lai, 2018).

Chen (2007) explored the effect of sample size ratios (e.g., 1:1; 1:2, and 1:4) on the sensitivity of fit indexes (e.g., $χ^2$, $Δ$CFI, $Δ$RMSEA, $Δ$SRMR, $Δ$Gamma hat, and $Δ$Me) and
found that the sample size ratio affected changes in the fit measures across metric, scalar, and residual invariance levels. Specifically, the changes in the fit measures were larger when sample sizes were equal than when sample sizes were unequal across three levels of invariance. In other words, under the condition of unequal sample size, it was less likely for MI tests to detect noninvariance. However, Chen (2007) considered only the model sizes of a single factor with 8 or 12 items. The findings might not generalize to various model size conditions.

Yoon & Lai (2008) examined the change of the fit measures (e.g., $\Delta\chi^2$, $\Delta$CFI, $\Delta$RMSEA, and $\Delta$SRMR) in MI testing with unequal sample sizes. Nine sample size conditions were monitored (e.g., 1:1, 1:2, 1:3, 1:4, 1:5, 1:7, 1:10, and 1:15) with the sample size for the first group set at 200. The study found that as the degree of imbalance increased, the mean of $\Delta$RMSEA and $\Delta$SRMR decreased and that of $\Delta$CFI increased to indicate a better fit of the misspecified invariance models. Overall, the power to detect the violation of invariance decreased as the sample size ratio became larger. The findings had practical implications. However, only one small model size (i.e., one factor with six indicators) was involved in this study. The results might not apply to MI tests with different model sizes involved.

The Impact of Model Size on Fit Measures in Measurement Invariance

Previous research has examined the effect of model sizes on the sensitivity of fit measures to detect MI. Cheung and Rensvold (2002) examined the sensitivity of three fit indices (e.g., $\Delta$CFI, $\Delta$Gamma, $\Delta$McDonald’s) on detecting MI by considering different model sizes. The number of factors (2 or 3) and the number of items per factor (3, 4, or 5)
were simulated. The study results showed that $\Delta$CFI, $\Delta$Gamma, $\Delta$McDonald’s were independent of model complexity.

French and Finch (2008) investigated the MI using $\Delta \chi^2$, $\Delta$CFI, and a combination of $\Delta \chi^2$ and $\Delta$CFI. Data were simulated from two- and four-factor models with the number of indicators per factor (three and six) considered. The study found that the higher power of $\Delta \chi^2$ was associated with a greater number of factors and indicators per factor. However, the power of $\Delta$CFI was higher with fewer factors.

Chen (2007) also conducted a simulation study to investigate the sensitivity of fit indices ($\Delta$CFI, $\Delta$RMSEA, and $\Delta$ SRMR) to detect a lack of MI for metric, scalar, and residual invariance with two model size conditions (e.g., a single factor with 8 or 12 indicators). The results of the study showed that the number of indicators did not have a significant impact on $\Delta$CFI, $\Delta$SRMR, or $\Delta$ Gammahat. The $\Delta$RMSEA was bigger in the 8-indicator models than that in the 12-indicator models.

Meade et al. (2008) conducted a simulation study with different numbers of factors (i.e., 2 and 4 factors) and indicators per factor (i.e., 4 and 8 items) involved. Findings showed that some of the fit indices (e.g., $\Delta$CFI, $\Delta$McDonald’s, $\Delta$IFI, $\Delta$RNI, $\Delta$Gamma, and $\Delta$SRMR) were sensitive to a lack of MI while being unaffected by the number of factors or the number of indicators.

Fan and Sivo (2009) examined several fit indices (e.g., $\Delta$RMSEA, $\Delta$NFI, $\Delta$TLI, $\Delta$CFI, $\Delta$Gamma, and $\Delta$McDonald’s) for mean structure invariance with different model sizes simulated (e.g., 2, 3, and 4 factors with 2, 4, and 6 indicators per factor). The study results showed that the indices were too sensitive to model complexity and should not be used for mean structure invariance testing.
A recent study by Cao and Liang (2022) examined the impact of the model size on the sensitivity of fit measures (e.g., CFI, RMSEA, AIC, BIC, SaBIC, LRT) to MI by simulating 6 levels of the model sizes (1 factor with 8 or 16 items, 2 factors and 4 factors with 4 or 8 items per factor). The study found that CFI and RMSEA failed to detect metric invariance when the model size was large (e.g., 6 factors with 8 items per factor). AIC and LRT were minimally affected by the model size.

Across the studies reviewed, findings from previous research were inconsistent due to the differences in model sizes. Besides, all these previous studies, except the study by Chen (2007), examined the impact of model size on the sensitivity of model fit statistics in MI testing without considering the potential combined impact of group size ratio. The findings from previous research (Chen, 2007; Yoon & Lai, 2018) indicated that group size ratio affected the performance of fit statistics in MI testing. However, no previous research has studied this combination. Therefore, the current study explored the impact of the model sizes in combination with the group size ratio on the sensitivity of changes on three model fit indices (i.e., ΔCFI, ΔRMSEA, and ΔSRMR) for establishing the metric and scalar MI.

**Method**

**Population Model**

**Number of groups**

Data were generated and analyzed with the MG-CFA model. Two groups were considered the most common number of groups involved in MI tests (70.6%) based on the Study 1 review results. Here, Group 1 was considered the reference group, and Group 2 was the focal group. Parameter values for Group 2 were manipulated to create different conditions for non-invariance.
Data characteristics

The data for the simulation were continuous and followed a multivariate normal distribution. This condition was used as criteria for the fit measures investigated were developed with normally distributed continuous data (Chen, 2007). Therefore, to make the results comparable, continuous data with normal distribution were used. The maximum likelihood (ML) estimator was used for data analysis (Kline, 2015). ML achieves optimal performance under this condition.

Number of Factors and Factor loadings.

The number of factors was fixed to two as it was one of the most commonly used factor structures in the previous simulation study (e.g., Cheung & Rensvold, 2002; Fan & Sivo, 2009; French & Finch, 2006; Meade et al., 2008). Factor loading values greater than 0.7 are suggested for structural equation modeling (Kline, 2015) because items with factor loadings greater than 0.7 are likely to have interpreting power (Lin & Cheng, 2020). Also, larger factor loading values (e.g., 0.8) generally yielded fewer problems with statistical power relative to weaker factor loadings (e.g., 0.5) (Wolf et al., 2013). Therefore, the factor loadings of all the items in Group 1 and all the metric invariant items in Group 2 were set at 0.8 in the population model.

Pattern of noninvariance.

A uniform pattern of noninvariance (i.e., one group has higher loadings on all items than the other group) was adopted (Chen, 2007; Chen, 2008; Meade & Lautenschlager, 2004), with lower loadings placed on non-invariant items in Group 2. This pattern of non-invariance indicated that the direction of lack of invariance was the same (Chen, 2008).
Item Intercepts

Item intercept (i.e., mean) values for all items in Group 1 and scalar invariant items in Group 2 were set at zero. These values were chosen because they have been used by previous researchers (e.g., Kim et al., 2012; Liang & Luo, 2020).

Item Residual Variances

Residual variances of all invariant items were defined as 1.0 minus the square of factor loadings, that is, 0.36. Those for non-invariant items were set based on the magnitude of the invariance.

Factor Means, Variances, and Correlation.

Factor means and variances were set to zero and one for both groups (Cao & Liang, 2022; Liang & Luo, 2020). The factor correlation value was fixed to be 0.5 to indicate a moderate correlation between factors (French & Finch, 2008).

Proportion of non-invariant items

The proportion of 20% non-invariant items, which indicated a low contamination situation, was used in the current study consistently with previous simulation research (e.g., French & Finch, 2008; Meade & Wright, 2012).

Magnitude of Invariance

The magnitude of cross-group differences was set to 0.4 for factor loadings and intercepts (e.g., Kim et al., 2012; Kim & Yoon, 2011; Shi et al., 2017; Thompson et al., 2021).

Sample Size

Wolf et al. (2013) recommended a minimum sample size range from 300 to 460 for structural equation modeling (SEM) while considering the number of indicators and
factors, the magnitude of factor loadings, path coefficients, and the amount of missing data. Therefore, the current study fixed the total sample size to 500. Table 4.1 summarizes all fixed conditions investigated.

**Design Factors**

**Model size**

The model size (i.e., the number of items on two factors) was manipulated in this study. The review study on CFA by Zhao (2015) showed that various model sizes were adopted in applied research, including models with 10 or less variables (15%), models with 11 to 40 variables (64%), and models with more than 40 items (21%). The results showed that small to relatively large models were involved in CFA models. A minimum of three observed variables for a single factor is needed for model estimation and identification purpose (Kline, 2015; Raykov & Marcoulides, 2012). Two or more indicators are needed for CFA models with two or more factors (Kline, 2015). Besides, more indicators per factor than the minimum number compensate for small sample size preserve statistical power (Wolf et al., 2013). Therefore, four model sizes were considered in the current study: small (i.e., 2 factors/10 indicators), medium (i.e., 2 factors/20 indicators), and large (i.e., 2 factors/30 indicators).

**Location of non-invariance**

The location of non-invariance affected the power detecting MI using $\Delta \chi^2$ (Sass et al., 2014). In the current study, MI was simulated to occur in two locations: factor loadings only or item intercepts only, as previous research did (e.g., Cao & Liang, 2022; Shi, 2017). Specifically, measurement non-invariance occurred in Group 2. For the factor loading non-invariance conditions, the factor loading(s) of the non-invariant indicator(s) in Group 2
were fixed by subtracting a constant (i.e., the magnitude of the noninvariance) from the factor loadings of the corresponding indicator(s) in the reference group (e.g., González-Romá et al., 2006; Kim & Yoon, 2011). For the intercept non-invariance conditions, the intercept(s) of the non-invariant indicator(s) in Group 2 were obtained by subtracting a constant from the intercepts of the corresponding indicator(s) in Group 1 (e.g., Cao Liang, 2022; Kim & Yoon, 2011).

**Group size ratio**

Study 1 results showed that 1:1 and 1:2 were two commonly used group size ratios. Fewer MI tests involved groups with group size ratios greater than 1:3. Therefore, three group size ratios (1:1, 1:2, and 1:4) were manipulated in the current study. These group size ratios were also examined by previous research (Chen, 2007). The total sample size of 500 was examined in the study, resulting in group sample sizes of 250:250, 167:333, and 100:400.

**Location of non-invariant indicators**

Conceptually, non-invariant items could exist on only one factor or both factors. Therefore, this condition specifies the location of the noninvariant items. Specifically, one condition had non-invariant items on only one factor and the other had non-invariant items on both factors. Table 4.2 summarizes the design factors manipulated.

In sum, 48 conditions were included in the design (3 model sizes x 2 measurement noninvariance locations x 3 group size ratio x 2 locations of non-invariant indicators). These factors were used for the non-invariant data generation.
Data Analysis

Data were generated and analyzed with Mplus software package (v. 8.6; Muthén & Muthén, 1998-2017) and summarized with the R software package (R Core Team, 2020). One thousand replications were generated for each simulation condition. Any non-convergence or improper solutions were removed from the study, and the number of replications increased until 1000 successful iterations were achieved.

In the configural invariance model, the factor structure was constrained to be the same across groups, with item intercepts, factor loadings, and residuals freely estimated. For model identification, the latent factor means and factor variances for both groups were fixed at zero and one. In the metric invariance model, the factor loadings were constrained to be equal across the two groups. Item intercepts and residuals were freely estimated. Factor means in both groups were fixed at zero and the factor variance in Group 1 was fixed at one but freely estimated in Group 2 for model identification purposes (i.e., to obtain a statistically equivalent but identified model) (Millsap & Yun-Tein, 2004; Muthén & Asparouhov, 2002).

In the scalar invariance model, factor loadings and item intercepts were constrained to be equal across groups. Item residuals were freely estimated. Also, factor means and variances for Group 1 were fixed to zero and one but freely estimated in Group 2.

Configural invariance was assumed in the current study. Metric invariance was tested by comparing the constrained metric invariance model to the nested, unconstrained configural invariance model. Scale invariance was tested by comparing the constrained scalar invariance model to the nest, unconstrained metric model. The current study used multiple model fit information criteria as reliance only on one model fit measure can result
in Type II errors (Yoon & Millsap, 2007). In other words, using only one model fit statistic may lead to the failure to detect violations of invariance.

Three commonly used model fit statistics in research and practice were investigated, including $\Delta$RMSEA, $\Delta$CFI, and $\Delta$SRMR (Cheung & Rensvold, 2002; Joo & Kim, 2019; Putnick & Bornstein, 2016). Chen (2007) suggested the criteria of $\Delta$CFI $\leq$ .010, $\Delta$RMSEA $\leq$ .015, and $\Delta$SRMR $\leq$ .30 for metric invariance and $\Delta$CFI $\leq$ .010, $\Delta$RMSEA $\leq$ .015, and $\Delta$SRMR $\leq$ 0.01 for scalar and residual invariance when the sample size is larger than 300 and the sample size is equal to two groups. However, as the current study involved unequal sample sizes and previous research found that the mean of SRMR decreased when the degree of imbalance increased (Yoon & Lai, 2018), cutoff values of $\Delta$CFI $\leq$ 0.01, $\Delta$RMSEA $\leq$ 0.015, and $\Delta$SRMR $\leq$ 0.01 were chosen as the criteria to determine power rates. Statistical power was used to assess the proportion of replications that correctly identified noninvariant models as noninvariant using the cutoff criteria for $\Delta$RMSEA, $\Delta$CFI, and $\Delta$SRMR. Power rates were evaluated against the cutoff values of $\beta \geq 0.80$ (Kline, 2015).

**Results**

**Convergence Rates**

Overall convergence rates were high across all simulated conditions. Convergence problems occurred only once, resulting in the lowest convergence rate of 99.9%.

**Metric Invariance Testing**

**Power of $\Delta$CFI**

As shown in Table 4.3 and Figure 4.1, the power of $\Delta$CFI for identifying non-invariance was affected by group ratio and the location of noninvariant items across the three model sizes. Regarding group ratio, the power of $\Delta$CFI remained stable across group
ratios of 1:1 and 1:2 (e.g., β of 0.961 vs. β of 0.962 in Model size 1 and location of non-invariant items on one factor condition). However, the power of ΔCFI decreased when the group ratio changed from 1:1 or 1:2 to 1:4 (e.g., β of 0.961 vs. β of 0.848 in Model size 1 and location of non-invariant items on one factor condition). However, the power of ΔCFI decreased when the group ratio changed from 1:1 or 1:2 to 1:4 (e.g., β of 0.961 vs. β of 0.848 in Model size 1 and location of non-invariant items on one factor condition).

Regarding the location of non-invariant items, the power of ΔCFI was stronger when the non-invariant items were located on two factors rather than only one factor (e.g., β of 0.851 vs. β of 0.960 in model size 3 and group size ratio of 1:4 condition). Overall, the ΔCFI power was the lowest when a group size ratio of 1:4 was involved and non-invariant items were located on only one factor, with all the power rates under 0.90 regardless of the model size. Regardless of the group size ratio or non-invariant items location, the ΔCFI power was the lowest for model size 2 (i.e., 2 factors with 20 indicators) (mean = 0.93). The mean ΔCFI power was the same (mean = 0.95) under model size 1 (2 factors/10 indicators) and model size 3 (2 factors/30 indicators) conditions. In general, ΔCFI power was adequate (i.e., power rates across conditions were above 0.80) for identifying metric non-invariance across all conditions.

**Power of ΔRMSEA**

As presented in Table 4.3 and Figure 4.2, the power of ΔRMSEA for identifying non-invariance was affected by group ratio and the location of noninvariant items across the three model sizes. Concerning group ratio, the power of ΔRMSEA did not show much change across group ratios of 1:1 and 1:2 (e.g., β of 0.999 vs. β of 0.994 in model size 1 and location of non-invariant items on one factor condition). However, the power of ΔRMSEA decreased when the group size ratio reached 1:4. For example, β dropped from 0.999 (1:1) to 0.969 (1:4) in model size 1 and location of non-invariant items on one factor.
condition. $\beta$ decreased from 0.949 (1:1) to 0.763 (1:4) in model size 2 and location of non-invariant items on one factor condition. Under the model size 3 and location of non-invariant items on one factor condition, $\beta$ dropped from 0.702 (1:1) to 0.221 (1:4). The results showed that $\Delta$RMSEA did not perform well for identifying non-invariance when the group size ratio was large, especially in larger model size.

Regarding location, $\Delta$RMSEA illustrated higher power when the non-invariant items occurred on both factors rather than only one factor (e.g., $\beta$ of 0.952 vs. $\beta$ of 0.983 in model size 2 and group size ratio of 1:2 condition). What was seen indicated that the number of non-invariant items per factor affects the power of $\Delta$RMSEA for detecting noninvariance.

Model size had the greatest impact on the power of $\Delta$RMSEA for identifying non-invariance. Overall, the power of $\Delta$RMSEA decreased as model sizes increased. Specifically, under model size 1 (2 factors/10 indicator) condition, the power of $\Delta$RMSEA was adequate for identifying the non-invariance (e.g., range of $\beta$ from 0.969 to 1.000). However, under model size 2 condition (2 factors/20 indicators), the $\Delta$RMSEA failed to identify the non-invariance under one condition (i.e., $\beta$ is 0.763 under the condition of 1:4 ratio and the location of non-invariant items on one factor). Under the model size 3 condition (2 factors/30 indicators), $\Delta$RMSEA performed acceptably to identify non-invariance only under two conditions (i.e., 1:1 or 1:2 and the location of non-invariant items on both factors). However, the power rates of $\Delta$RMSEA were below 0.80 under rest conditions with model size 3. In general, $\Delta$RMSEA was adequate for identifying non-invariance when the model size was small.
Power of ΔSRMR

As exhibited in Table 4.3 and Figure 4.3, the power of ΔSRMR for identifying metric non-invariance was not stable across conditions. Regarding group size ratio, the power of ΔSRMR reached the lowest when a group size ratio of 1:2 was involved. For example, β dropped from 0.946 to 0.233 under the condition of model size 2 and the location of non-invariant items on one factor.

The location of the non-invariant items did not make much difference in the power of ΔSRMR for identifying the non-invariance. For example, under model size 3 and group size ratio of 1:1 condition, β was similar when noninvariant items were located on one factor (0.994) or two factors (0.992).

Model size had the greatest impact on the power of ΔSRMR. When the model size was small (i.e., 2 factors/10 indicators), the power of ΔSRMR was adequate only under the condition of a group size ratio of 1:4, with β values under 0.80 under the rest of the conditions. When the model size was medium (i.e., 2 factors/20 indicators) or large (i.e., 2 factors/30 indicators), the power of ΔSRMR failed to identify the non-invariance under the condition of group size ratio of 1:2.

Scalar Invariance Testing

Power of ΔCFI

As shown in Table 4.4 and Figure 4.4, the power of ΔCFI was similar under the group size ratio of 1:1 and 1:2; (e.g., β of 0.992 vs. β of 0.968 in Model size 1 and location of non-invariant items on one factor condition) across the three model sizes. However, the power of ΔCFI failed to identify scalar noninvariance when the group size ratio was 1:4, with β values under 0.80 regardless of model size or non-invariant item location.
The location of non-invariant items affected the power of ΔCFI for identifying the non-invariance, with β values consistently higher when non-invariant items were located on two factors than on one factor across other conditions. For example, β increased from 0.770 to 0.939 under model size 1 and the group size ratio of 1:4 condition. Specifically, the ΔCFI failed to identify the non-invariance when non-invariant items were located on one factor and the group size ratio was 1:4, with all the β values under 0.80.

Model size did not make much difference when estimating the ΔCFI power. The ΔCFI power was the highest for model size 1 (2 factors/10 indicators)(mean=0.94). The mean ΔCFI power was the same (mean=0.90) under model size 2 (2 factors/20 indicators) and model size 3 (2 factors/30 indicators) conditions.

Power of ΔRMSEA

As exhibited in Table 4.4 and Figure 4.5, the power of ΔRMSEA was stable across group ratios of 1:1 and 1:2 (e.g., β of 0.991 vs. β of 1.000 in Model size 1 and location of non-invariant items on one factor condition). However, when the group size ratio changed from 1:1 or 1:2 to 1:4, the power of ΔRMSEA decreased. For example, β dropped from 0.991 (1:1) to 0.907 (1:4) in model size 1 and location of non-invariant items on one factor condition. β decreased from 0.954 (1:1) to 0.601 (1:4) in model size 2 and location of non-invariant items on one factor condition. Under the model size 3 and location of non-invariant items on one factor condition, β dropped from 0.757 (1:1) to 0.132 (1:4).

The power of ΔRMSEA was higher when the non-invariant items spread out on both factors rather than only one factor across all other conditions (e.g., β of 0.907 vs. β of 0.993 in model size 2 and group size ratio of 1:2 condition), indicating that ΔRMSEA was
more likely to detect scalar non-invariance when non-invariant items located on one factor rather than two factors.

The power of ΔRMSEA for identifying non-invariance was affected by the model sizes. Overall, as model sizes increased, the power of ΔRMSEA decreased. The power of ΔRMSEA was adequate for identifying the non-invariance when a small model size (2 factors/10 indicators) was involved. Under the medium model size condition (2 factors/20 indicators), the ΔRMSEA failed to identify the non-invariance under the condition of 1:4 ratio and the location of non-invariant items on one factor (i.e., β was 0.601). Under the model size 3 condition (2 factors/30 indicators), ΔRMSEA was adequate for identifying non-invariance only under two conditions (i.e., 1:1 or 1:2 and the location of non-invariant items on both factors). Overall, the ΔRMSEA functioned well when a small model size was involved.

**Power of ΔSRMR**

ΔSRMR failed to identify scalar non-invariance across any conditions, indicating that ΔSRMR had extremely low sensitivity to the scalar non-invariance.

**Discussion**

While different model sizes and group size ratios were involved in empirical research on MI testing, the suggested criteria of fit measures were commonly used by researchers for evaluating MI regardless of the model sizes or group size ratios. No prior research has explored the performance of fit measures when different model sizes and group size ratios were involved in MI testing. Therefore, this study examined the impact of group size ratio in combination with model size on the sensitivity of three fit measures (i.e., ΔCFI, ΔRMSEA, and ΔSRMR) to detect metric and scalar non-invariance.
Using a simulation study with an MG-CFA population model, the characteristics of the model were varied to approximated conditions encountered with empirical studies. Model size, group size ratio, location of noninvariance, and location of noninvariant items were manipulated to investigate the effect of group size ratio and model size on the sensitivity of ΔCFI, ΔRMSEA, and ΔSRMR in detecting metric or scalar noninvariance. The findings of this study can provide recommendations for researchers applying these three commonly used fit measures to evaluate MI testing under different group size rations and model sizes.

**Metric Invariance Testing**

The sensitivity of CFI in detecting metric noninvariance was affected by group size ratio and the location of non-invariant items. The sensitivity of ΔCFI did not change much between the group size ratio of 1:1 and 1:2. However, the sensitivity of CFI deteriorated when the group size ratio changed from 1:1 or 1:2 to 1:4. The finding corresponded with the previous research results that unequal sample sizes are more likely to reduce power in detecting noninvariance than equal sample size condition (Chen, 2007; Kaplan & George, 1995), especially when the sample size in two groups were quite different (Chen, 2007; Yoon& Lai, 2018).

The current study added additional evidence that the sensitivity of ΔCFI was similar across group size ratios of 1:1 and 1:2. ΔCFI performed better when non-invariant items were spread equally into two factors than on one factor. This finding implied that it would be easier for ΔCFI to identify non-invariance if fewer non-invariant items load on a factor. ΔCFI performed similarly across different model sizes, which was in line with the previous finding that the variance in the sensitivity of ΔCFI explained by the number of
items per factor was small (Cao & Liang, 2022). Overall, ΔCFI performed well in identifying metric noninvariance across all the conditions, which was in line with the findings by Meade et al. (2008) that ΔCFI is one of the best indices to report while conducting MI testing.

Similar to ΔCFI, ΔRMSEA was also affected by the group size ratio and the location of non-invariant items in the same way. The sensitivity of ΔRMSEA decreased when the group size ratio changed from 1:1 or 1:2 to 1:4. The finding was supported by a previous study that unbalanced sample size increased the ΔRMSEA in support of the metric invariance model (Yoon & Lai, 2018). Unlike ΔCFI, ΔRMSEA was affected by the model size. When the model size increased, the sensitivity of ΔRMSEA in detecting metric non-invariance decreased. ΔRMSEA performed well with a small model size; however, the index failed to detect metric noninvariance under four out of six conditions in a large model size. These findings indicated that ΔRMSEA was less sensitive to the model misspecification when model size was large. The findings here were similar to results found in previous simulation studies that ΔRMSEA failed to detect metric noninvariance in large model sizes (e.g., Cao & Liang, 2022; Fan & Sivo, 2007; Kenny & McCoach, 2003). Specifically, when the model size was large, the sensitivity of ΔRMSEA failed to detect the non-invariance when non-invariant items were located on only one factor, regardless of the group size ratio. This might be due to the reason that when the percentage of noninvariance passed a certain threshold, the changes in the goodness of fit indices did not detect the noninvariance (Chen, 2007). When the number of items increased in larger model sizes, the number of noninvariant items also increased (i.e., 6 non-invariant items located on one factor in the 2 factors/30 indicators model).
Compared to ΔRMSEA and Δ CFI, the sensitivity of ΔSRMR was less affected by the location of the non-invariant items. The sensitivity of ΔSRMR remained similar regardless of whether the non-invariant items were located on one or two factors. ΔSRMR was more likely to identify metric non-invariance in larger model sizes. In a small model size, ΔSRMR performed well only when the group size ratio was 1:4. In the larger model sizes, ΔSRMR failed to detect metric non-invariance when the group size ratio was 1:2. The inconsistent performance showed that the group size ratio interacted with model size to have an impact of the sensitivity of ΔSRMR.

Scalar Invariance Testing

For scalar noninvariance testing, the sensitivity of the ΔCFI was unaffected by the model size. The findings corresponded to previous research, which also found that the sensitivity of CFI for detecting scalar noninvariance was not affected by the model size, especially for detecting the magnitude of scalar noninvariance of 0.30 (Cao & Liang, 2022). However, it failed to detect the non-invariance under one condition (e.g., a 1:4 ratio in the combination of the location of non-invariant items on only one factor).

The sensitivity of the ΔRMSEA to scalar noninvariance was the same as to the metric noninvariance, indicating that ΔRMSEA performed the same way for detecting scalar or metric noninvariance. ΔSRMR failed to detect the scalar non-invariance under any condition. The finding corresponded with the previous research results (Yoon & Lai, 2018). This might be due to the reason that SRMR measures the average distance between a sample covariance matrix and its model-implied matrix without considering the discrepancy arising from the mean structure. Therefore, it is less likely for SRMR to detect the mean structure (Yoon & Lai, 2018). Overall, as ΔSRMR performed poorly in detecting
scalar noninvariance in all conditions, SRMR is not recommended for detecting scalar non-invariance even if metric invariance holds.

**Practical Implications**

Based on the findings, some recommendations can be made for applied researchers interested in MI testing with the MG-CFA approach. When researchers conduct scalar MI testing that involves a large unbalanced group size (e.g., 1:4), especially when the non-invariant items are located on only one factor, the suggested cutoff criteria of the $\Delta CFI \leq 0.01$ should not be used for evaluating scalar invariance due to insensitivity to the combination of the conditions (i.e., group size ratio of 1:4 and location of non-invariant items on one factor). However, using a $\Delta CFI \leq 0.01$ was adequate for detecting metric invariance. The suggested cutoff criteria of the $\Delta RMSEA \leq 0.015$ should be avoided for testing either metric or scalar invariance in large model size (e.g., 2 factors/30 indicators), especially when the group size ratio was more unbalanced (e.g., 1:4) and non-invariant items are located on only one factor. However, the criteria could be applied to testing metric and scalar invariance when a small model size is tested (e.g., 2 factors/10 indicators) because the power of $\Delta SRMR \leq 0.01$ was adequate for detecting scalar noninvariance across all the conditions in a small model size. The suggested cutoff criteria of the $\Delta SRMR \leq 0.01$ should not be used for detecting scalar invariance. Besides, it was not recommended to be used when a group ratio of 1:2 is involved, regardless of the model size or location of the non-invariant items.

**Limitations and Future Directions**

As with other simulation studies, the current study manipulated only a limited number of conditions and has limitations. First, this study considered only continuous
observed variables with a normal distribution. However, as variables characterized by an ordinal level of measurement are common in empirical research (Flora & Curran, 2004), future research may use ordinal variables to explore the performance of fit measures in MI testing with different group size ratios combined with varying model sizes.

Second, only two groups were used for the invariance testing. Study 1 review results showed that MI tests often involved more than two groups. Rutkowski and Svetina (2017) indicated that $\Delta$CFI decreased and $\Delta$RMSEA increased as the number of groups increased from 10 to 20 for metric MI testing. It remains unknown about the sensitivity of fit measures to MI with a large number of groups combined with varying group size ratios. Future research might explore the effects of group size ratio on testing MI with more than two groups.

Third, the present study considered only one type of proportion of non-invariance (20%). As previous research found that the proportion of non-invariant items affects the performance of the fit measures in detecting MI (Chen, 2007; Rutkowski & Svetina, 2017), future research can add more proportion levels, such as 25%, 50% or 75% (Chen, 2007) to examine whether the proportion of non-invariance in combination with group size ratio and model size has an impact on the sensitivity of fit measures in detecting MI.

Finally, the factor mean and variance was equal across the two groups compared in this study. Future research may explore the same topic under the condition of unequal factor means and variances across groups.

In summary, the present study evaluated the impact of the model size and group size ratio on the sensitivity of three commonly used fit measures ($\Delta$CFI, $\Delta$RMSEA, and $\Delta$SRMR) for detecting the lack of the MI with MG-CFA approach. The findings of the
current study suggested that $\Delta \text{CFI} \leq 0.01$ criterion was adequate for detecting metric noninvariance. However, it failed to detect scalar noninvariance when a group size ratio of 1:4 and the location of non-invariant items on one factor were involved. $\Delta \text{RMSEA} \leq 0.015$ functioned well in detecting metric or scalar noninvariance when a small model size was involved (e.g., 2 factors/10 indicators). However, $\Delta \text{RMSEA} \leq 0.015$ did not perform well in detecting either metric or scalar invariance with a large model size (e.g., 2 factors/30 indicators), especially when the group size ratio was 1:4 and the location of non-invariant items on one factor was involved. $\Delta \text{SRMR} \leq 0.010$ failed to detect metric invariance when a group ratio of 1:2 was involved. $\Delta \text{SRMR} \leq 0.010$ was not adequate to detect scalar invariance. The findings of the current study added to the body of research in MI testing literature and provided suggestions for applied researchers interested in MI testing under an MG-CFA approach.

References


Gregorich, S. E. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework. *Medical Care, 44*(11 Suppl 3), S78. [https://doi.org/10.1097/01.mlr.0000245454.12228.8f](https://doi.org/10.1097/01.mlr.0000245454.12228.8f)


Table 4.1 Fixed parameters in the population model

<table>
<thead>
<tr>
<th>Parameter controlled</th>
<th>Population value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group number</td>
<td>2</td>
</tr>
<tr>
<td>Number of factors</td>
<td>2</td>
</tr>
<tr>
<td>Item loading</td>
<td>0.8</td>
</tr>
<tr>
<td>Item intercept</td>
<td>0</td>
</tr>
<tr>
<td>Item residual</td>
<td>0.36</td>
</tr>
<tr>
<td>Factor mean</td>
<td>0</td>
</tr>
<tr>
<td>Factor variance</td>
<td>1</td>
</tr>
<tr>
<td>factor correlation</td>
<td>0.5</td>
</tr>
<tr>
<td>Proportion of non-invariant items</td>
<td>20%</td>
</tr>
<tr>
<td>Magnitude of invariance</td>
<td>0.4 for factor loading and intercepts</td>
</tr>
<tr>
<td>Item distribution</td>
<td>multivariate normal distribution</td>
</tr>
<tr>
<td>Sample size</td>
<td>500</td>
</tr>
</tbody>
</table>

Table 4.2 Simulation design factors and manipulations

<table>
<thead>
<tr>
<th>Design Factor</th>
<th>Manipulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model sizes (number of indicators)</td>
<td>10, 20, 30</td>
</tr>
<tr>
<td>group size ratio</td>
<td>1:1, 1:2, 1:4</td>
</tr>
<tr>
<td>Location of non-invariance</td>
<td>factor loadings only or intercept only</td>
</tr>
<tr>
<td>Location of non-invariant indicators</td>
<td>one factor/ two factors</td>
</tr>
</tbody>
</table>
Table 4.3 Power rates of ΔCFI, ΔRMSEA, and ΔSRMR for detecting metric noninvariance

<table>
<thead>
<tr>
<th>Model Size</th>
<th>Group size ratio</th>
<th>Location of non-invariant items</th>
<th>ΔCFI</th>
<th>ΔRMSEA</th>
<th>ΔSRMR</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 factors /10 items</td>
<td>1:1</td>
<td>1 factor</td>
<td>0.961</td>
<td>0.999</td>
<td>0.441</td>
</tr>
<tr>
<td>2 factors/10 items</td>
<td>1:1</td>
<td>2 factors</td>
<td>0.996</td>
<td>1.000</td>
<td>0.246</td>
</tr>
<tr>
<td>2 factors/10 items</td>
<td>1:2</td>
<td>1 factor</td>
<td>0.962</td>
<td>0.994</td>
<td>0.019</td>
</tr>
<tr>
<td>2 factors/10 items</td>
<td>1:2</td>
<td>2 factors</td>
<td>0.991</td>
<td>1.000</td>
<td>0.012</td>
</tr>
<tr>
<td>2 factors/10 items</td>
<td>1:4</td>
<td>1 factor</td>
<td>0.848</td>
<td>0.969</td>
<td>1.000</td>
</tr>
<tr>
<td>2 factors/10 items</td>
<td>1:4</td>
<td>2 factors</td>
<td>0.937</td>
<td>0.991</td>
<td>1.000</td>
</tr>
<tr>
<td>2 factors/20 items</td>
<td>1:1</td>
<td>1 factor</td>
<td>0.948</td>
<td>0.949</td>
<td>0.946</td>
</tr>
<tr>
<td>2 factors/20 items</td>
<td>1:1</td>
<td>2 factors</td>
<td>0.989</td>
<td>0.990</td>
<td>0.921</td>
</tr>
<tr>
<td>2 factors/20 items</td>
<td>1:2</td>
<td>1 factor</td>
<td>0.939</td>
<td>0.952</td>
<td>0.233</td>
</tr>
<tr>
<td>2 factors/20 items</td>
<td>1:2</td>
<td>2 factors</td>
<td>0.984</td>
<td>0.983</td>
<td>0.152</td>
</tr>
<tr>
<td>2 factors/20 items</td>
<td>1:4</td>
<td>1 factor</td>
<td>0.820</td>
<td>0.763</td>
<td>1.000</td>
</tr>
<tr>
<td>2 factors/20 items</td>
<td>1:4</td>
<td>2 factors</td>
<td>0.926</td>
<td>0.876</td>
<td>1.000</td>
</tr>
<tr>
<td>2 factors/30 items</td>
<td>1:1</td>
<td>1 factor</td>
<td>0.944</td>
<td>0.702</td>
<td>0.994</td>
</tr>
<tr>
<td>2 factors/30 items</td>
<td>1:1</td>
<td>2 factors</td>
<td>0.988</td>
<td>0.847</td>
<td>0.992</td>
</tr>
<tr>
<td>2 factors/30 items</td>
<td>1:2</td>
<td>1 factor</td>
<td>0.963</td>
<td>0.640</td>
<td>0.516</td>
</tr>
<tr>
<td>2 factors/30 items</td>
<td>1:2</td>
<td>2 factors</td>
<td>0.995</td>
<td>0.807</td>
<td>0.479</td>
</tr>
<tr>
<td>2 factors/30 items</td>
<td>1:4</td>
<td>1 factor</td>
<td>0.851</td>
<td>0.221</td>
<td>1.000</td>
</tr>
<tr>
<td>2 factors/30 items</td>
<td>1:4</td>
<td>2 factors</td>
<td>0.960</td>
<td>0.320</td>
<td>1.000</td>
</tr>
</tbody>
</table>
Table 4.4 Power rates of \( \Delta \text{CFI} \), \( \Delta \text{RMSEA} \), and \( \Delta \text{SRMR} \) for detecting scalar noninvariance

<table>
<thead>
<tr>
<th>Model size</th>
<th>Group size ratio</th>
<th>Location of non-invariant items</th>
<th>( \Delta \text{CFI} )</th>
<th>( \Delta \text{RMSEA} )</th>
<th>( \Delta \text{SRMR} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 factors /10 items</td>
<td>1:1</td>
<td>1 factor</td>
<td>0.992</td>
<td>0.991</td>
<td>0.005</td>
</tr>
<tr>
<td>2 factors/10 items</td>
<td>1:1</td>
<td>2 factors</td>
<td>1.000</td>
<td>1.000</td>
<td>0.049</td>
</tr>
<tr>
<td>2 factors/10 items</td>
<td>1:2</td>
<td>1 factor</td>
<td>0.968</td>
<td>0.981</td>
<td>0.000</td>
</tr>
<tr>
<td>2 factors/10 items</td>
<td>1:2</td>
<td>2 factors</td>
<td>1.000</td>
<td>0.997</td>
<td>0.000</td>
</tr>
<tr>
<td>2 factors/10 items</td>
<td>1:4</td>
<td>1 factor</td>
<td>0.770</td>
<td>0.907</td>
<td>0.000</td>
</tr>
<tr>
<td>2 factors/10 items</td>
<td>1:4</td>
<td>2 factors</td>
<td>0.939</td>
<td>0.984</td>
<td>0.000</td>
</tr>
<tr>
<td>2 factors/20 items</td>
<td>1:1</td>
<td>1 factor</td>
<td>0.990</td>
<td>0.954</td>
<td>0.0000</td>
</tr>
<tr>
<td>2 factors/20 items</td>
<td>1:1</td>
<td>2 factors</td>
<td>0.999</td>
<td>0.999</td>
<td>0.000</td>
</tr>
<tr>
<td>2 factors/20 items</td>
<td>1:2</td>
<td>1 factor</td>
<td>0.958</td>
<td>0.907</td>
<td>0.000</td>
</tr>
<tr>
<td>2 factors/20 items</td>
<td>1:2</td>
<td>2 factors</td>
<td>0.998</td>
<td>0.993</td>
<td>0.000</td>
</tr>
<tr>
<td>2 factors/20 items</td>
<td>1:4</td>
<td>1 factor</td>
<td>0.591</td>
<td>0.601</td>
<td>0.000</td>
</tr>
<tr>
<td>2 factors/20 items</td>
<td>1:4</td>
<td>2 factors</td>
<td>0.917</td>
<td>0.876</td>
<td>0.000</td>
</tr>
<tr>
<td>2 factors/30 items</td>
<td>1:1</td>
<td>1 factor</td>
<td>0.995</td>
<td>0.757</td>
<td>0.000</td>
</tr>
<tr>
<td>2 factors/30 items</td>
<td>1:1</td>
<td>2 factors</td>
<td>1.000</td>
<td>0.979</td>
<td>0.000</td>
</tr>
<tr>
<td>2 factors/30 items</td>
<td>1:2</td>
<td>1 factor</td>
<td>0.978</td>
<td>0.574</td>
<td>0.000</td>
</tr>
<tr>
<td>2 factors/30 items</td>
<td>1:2</td>
<td>2 factors</td>
<td>1.000</td>
<td>0.927</td>
<td>0.000</td>
</tr>
<tr>
<td>2 factors/30 items</td>
<td>1:4</td>
<td>1 factor</td>
<td>0.488</td>
<td>0.132</td>
<td>0.000</td>
</tr>
<tr>
<td>2 factors/30 items</td>
<td>1:4</td>
<td>2 factors</td>
<td>0.936</td>
<td>0.349</td>
<td>0.000</td>
</tr>
</tbody>
</table>
Figure 4.1 Power rates of $\Delta$CFI for detecting metric noninvariance
Figure 4.2 Power rates of ΔRMSEA for detecting metric noninvariance
Figure 4.3. Power rates of $\Delta$SRMR for detecting metric noninvariance
Figure 4.4. Power Rates of ΔCFI for detecting scalar noninvariance
Figure 4.5. Power rates of ΔRMSEA for detecting scalar noninvariance.
CONCLUSION

While MI has been widely applied due to the development of statistical methods and techniques (Bauder, 2017), it is necessary to explore the practice of MI from different perspectives. The compilation of studies included in this dissertation examined the MI with MG-CFA approach from theoretical, empirical, and methodological perspectives. The studies provided implications for researchers regarding how to apply MI with MG-CFA in an appropriate way.

Previous research examined the MI conducted in different disciplines (Schmitt & Kuljanin, 2008; Vanderberg and Lance, 2000). However, no previous study examined the practice of MI in education with a systematic review method. Therefore, Study 1 conducted a systematic review of the application of MI with the MG-CFA approach in education.

Study 1 reviewed 63 papers on MI conducted with the MG-CFA approach. The papers were published in 51 education-related journals between January 2022 and June 2022. This systematic review summarized the MI reporting practices from different perspectives (e.g., Data screening, Estimation methods and fit indices, measurement invariance testing), identified issues and areas for further improvement, and provided guidelines for appropriate MI practices to researchers. Various issues were involved in the MI application. Issues related to data screening included inadequate sample sizes, missing information about item levels, the lack of reporting on missing data and how missing data were dealt with, lack of the discussion on the reliability of the instruments,
and the ignorance of the multilevel structure of the data. With regards to estimation methods and fit indices, the relevant issues included the lack of reporting of the estimator used, inconsistent use of the estimators across data types (i.e., continuous or categorical), lack of fitting CFA models to each group prior to MI tests, choice of few numbers of model fit measures, and lack of the local model fit reporting. Issues related to MI testing included MI testing only at less strict levels, few practices of partial invariance tests, and inconsistent choice of the fit measures for evaluating MI. In addition to showcasing the limitations of the MI practices conducted by researchers in the field of education, this review study also provided corresponding guidelines to deal with the relevant issues based on the theories put forward by previous research. Overall, the current study has theoretical implications for researchers conducting MI with the MG-CFA approach. The guidelines put forward in the study will help researchers to follow a step-by-step procedure for the appropriate application of MI. Besides, the issues identified in this study shed light on the topics to explore in Study 2 and Study 3.

One of the issues identified in Study 1 was that a small percentage (28.8%) of the MI tests were conducted with categorical data (i.e., data with item levels ranging from 2 to 4). The results implied that the practice of MI with categorical data was not common in education. Among these studies on MI tests with categorical data, the choice of estimator was inconsistent. Around 36.0% of the studies used ML-related estimators, and 64.0% used WLSMV-related ones. The findings showed that the practicing issues existed in the MI application. Starting from this point, Study 2 conducted an empirical study on applying MI with categorical data as a demonstration to applied researchers interested in MI testing with categorical data.
The Strengths and Difficulties Questionnaire (SDQ) is a brief screening instrument used to assess the behavioral attributes of children aged from 4 to 17 (Goodman, 2001). The previous research investigated the measurement invariance across different groups (DeVries et al., 2017; Gomez & Stavropoulos, 2020; Janitza et al., 2020; Liang et al., 2019; Murray et al., 2021). No previous study was conducted to examine the cultural MI of SDQ across China and the U.S. within elementary school contexts. Given the increasing educational exchange between China and the U.S., researchers are more interested in understanding the differences in mental health across the two cultures. Therefore, Study 2 examined the MI of the teacher-reported Strengths and Difficulties Questionnaire (SDQ) across the United States and China. A sample of elementary school students in the U.S. (N=639) and China (N=666) was used. The MI tests in Study 2 started with Data screening illustration, including software used (Mplus 8.4), group number (2), group size ratio (approximately 1:1), data type (categorical items with 3 levels), missing (no missing), reliability (Cronbach’s alpha reliability coefficients were reported for each group sample and total sample), data structure (Multilevel with students nested within teachers), multi-level structure considered (considered clustering design effects). Then, MI tested proceeded with CFA model fitting demonstration, including estimator (WLSMV to accommodate the categorical nature of the data), CFA model specifying (CFA with each group sample), model fit statistics (Chi-square statistics, CFI, RMSEA, SRMR), loading size reported (yes), factor loading values (greater than 0.62). Finally, the study ended with MI testing, including levels of MI tested (configural, metric, scalar, and residual invariance), fit measures used for evaluating MI (Δ χ^2, ΔCFI, ΔRMSEA, and ΔSRMR), criteria for fit measures (ΔCFI≤0.01, ΔRMSEA≤0.015, and
ΔSRMR ≤ 0.03 for the factor loadings invariance, and ΔCFI ≤ 0.01, ΔRMSEA ≤ 0.015, and ΔSRMR ≤ 0.01 for the item intercepts/ thresholds and residual invariance given sample size > 300), the invariance levels established (full configural, metric, scalar, and residual invariance).

Overall, Study 2 used an empirical example to showcase to researchers how to conduct an MI test with categorical data in an appropriate way. This study applied the guidelines developed by Study 1 and served as an extension and supplement to Study 1. The study has practice implications for researchers regarding how to adopt an MG-CFA approach for MI testing.

Another issue identified in Study 1 was that the suggested cutoff criteria of the fit measures (e.g., ΔCFI ≤ 0.010, ΔRMSEA ≤ 0.015, and ΔSRMR ≤ 0.30) were commonly used by researchers regardless of the model sizes or group ratio involved in MI testing. Study 1 review results showed that the sample size ratios ranged from 1:1 to 1:39. The number of factors ranged from 1 to 10 and that for item numbers was from 3 to 87. Previous research explored the impact of model size (e.g., Chen, 2007; Yoon & Lai, 2018) or group size ratio (e.g., Cheung & Renvold, 2002; Fan & Sivo, 2009; French & Finch, 2006; Meade et al., 2008) on the performance of fit measures in detecting MI. However, the impact of model size in combination with the group size ratio remained unknown. Therefore, Study 3 examined the impact of model size and group size ratio on the sensitivity of fit measures (ΔCFI, ΔRMSEA, and ΔSRMR) to detecting metric or scalar noninvariance.

Study 3 conducted a simulation study with an MG-CFA model. Parameter values were fixed in population model, including group number (2), factor number (2), item
loading for invariant items (0.80), Item intercept/residual for invariant items (0/0.36), factor mean/variance (0/1), factor correlation (0.50), the magnitude of invariance (0.4 for both factor loadings and intercepts), item distribution (multivariate normal distribution), and sample size (500). Different design factors were involved in the study, including models size (i.e., 2 factors/10 indicators, 2 factors/20 indicators, 2 factors/30 indicators), group size ratio (i.e., 1:1, 1:2, 1:4), location of non-invariant items (i.e., one factor or two factors), location of noninvariance (i.e., metric invariance only or scalar invariance only). This study found that both model size and group size ratio affected the sensitivity of fit measures to detecting metric or scalar noninvariance in different ways. Specifically, ΔCFI ≤ 0.01 criterion was sensitive to detecting metric noninvariance but not scalar noninvariance, especially under the combination of a group size ratio of 1:4 and location of non-invariant items on only one factor. ΔRMSEA ≤ 0.015 was sensitive to detecting both metric and scalar noninvariance in a small model size. However, it failed to detect either metric or scalar noninvariance under the combined condition of a group size ratio of 1:4 and the location of non-invariant items on one factor. ΔSRMR ≤ 0.010 failed to detect metric invariance when group ratio of 1:2 was involved. ΔSRMR ≤ 0.010 was not adequate to detect scalar invariance in any condition. The findings of Study 3 offered methodological implication for applied researchers regarding applying different fit measures to evaluating MI. It also served as an extension to Study 1 with regards to how to solve the issues with MI testing from a methodological perspective.

The three manuscripts in this dissertation contributed to the literature of MI with MG-CFA approach from different perspectives. Study 1 explored the MI from a theoretical perspective by conducting a systematic review study on MI practices. The
issues identified in this study helped raise researchers’ awareness of the inappropriate MI practices in their application and provided guidelines to researchers regarding how to implement MI in an appropriate way. This study served as the starting point of the series of studies in this dissertation and provided the theoretical foundation for the following Study 2 and Study 3. Study 2 gave an empirical example of how to conduct an MI test with categorical data. This study arose from one of the issues identified in Study 1 that the examination of MI with categorical data was not common. The demonstration gave researchers practical step-by-step guidance regarding how to implement MI tests. Study 3 also originated from another issue identified in Study 1 that researchers used suggested fit measure criteria regardless of the model sizes or group size ratios involved in their MI testing. Given the issues, this study examined the impact of the model size and group size ratios on the sensitivity of fit measures to detecting metric and scalar invariance. The study found that model size and group size ratio worked together to have an impact on the sensitivity of fit measures in detecting metric and scalar noninvariance, especially under the condition of large model size and group size ratio. The findings from this study provided researchers with methodological implications regarding using fit measures to evaluate MI. Overall, this series of studies explored MI from different perspectives, giving researchers a deeper understanding of MI with the MG-CFA approach theoretically, empirically, and methodologically.

References


Questionnaire. *Journal of pediatric psychology, 46*(10), 1249-1257. https://doi.org/10.31234/osf.io/u89rb


Bibliography


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](https://doi.org/10.1207/s15328007sem0902_5)


determination of measurement invariance. *Structural Equation Modeling, 13*(3),

the invariant referent sets. *Structural Equation Modeling: A Multidisciplinary Journal, 15*(1), 96-113. [https://doi.org/10.1080/10705510701758349](https://doi.org/10.1080/10705510701758349)


Gao, X., Shi, W., Zhai, Y., He, L., & Shi, X. (2013). Results of the parent-rated strengths
and difficulties questionnaire in 22,108 primary school students from 8 provinces

& Blumberg, S. J. (2019). Prevalence and treatment of depression, anxiety, and
[https://doi.org/10.1016/j.jpeds.2018.09.021](https://doi.org/10.1016/j.jpeds.2018.09.021)

Gignac, G. E. (2014). On the inappropriateness of using items to calculate total scale
score reliability via coefficient alpha for multidimensional scales. *European Journal of Psychological Assessment.*
[https://doi.org/10.1027/1015-5759/a000181](https://doi.org/10.1027/1015-5759/a000181)

[https://doi.org/10.1016/j.jsp.2006.05.005](https://doi.org/10.1016/j.jsp.2006.05.005)

Gomez, R., & Stavropoulos, V. (2020). Malaysian parent ratings of the strengths and
difficulties questionnaire: factor structure and measurement invariance across
[https://doi.org/10.1177/1073191118787284](https://doi.org/10.1177/1073191118787284)

of the mean and covariance structure analysis model for detecting differential
[https://doi.org/10.1207/s15327906mbr4101_3](https://doi.org/10.1207/s15327906mbr4101_3)

González, P., Nuñez, A., Merz, E., Brintz, C., Weitzman, O., Navas, E. L., ... & Gallo, L.
C. (2017). Measurement properties of the Center for Epidemiologic Studies
[https://doi.org/10.1037/pas0000330](https://doi.org/10.1037/pas0000330)


Gregorich, S. E. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the
https://doi.org/10.1097/01.mlr.0000245454.12228.8f

https://doi.org/10.1027/1015-7579/a000450

https://doi.org/10.3389/fpsyg.2014.00980


https://doi.org/10.1177/1073191116681493

https://doi.org/10.1016/j.appdev.2016.02.003

https://doi.org/10.1177/019874291303800305

https://doi.org/10.1007/s10802-012-9696-6


https://doi.org/10.1016/j.jvb.2010.10.003

https://doi.org/10.1080/10705519909540118


https://doi.org/10.1016/j.jad.2019.06.049

https://doi.org/10.1080/10705511.2019.1649983

https://doi.org/10.1016/j.jad.2019.06.049

https://doi.org/10.7326/0003-4819-151-4-200908180-00136


https://doi.org/10.1016/j.comppsych.2013.01.002

https://doi.org/10.1037/met0000075

https://doi.org/10.1177/1069397120915454

[https://doi.org/10.1207/s15327906mbr3801_5](https://doi.org/10.1207/s15327906mbr3801_5)

[https://doi.org/10.1037/1082-989x.9.3.275.sup](https://doi.org/10.1037/1082-989x.9.3.275.sup)

[https://doi.org/10.1080/00223891.2017.1281286](https://doi.org/10.1080/00223891.2017.1281286)

[https://doi.org/10.1080/10705510701575461](https://doi.org/10.1080/10705510701575461)

[https://doi.org/10.1037/0021-9010.93.3.568](https://doi.org/10.1037/0021-9010.93.3.568)

[https://doi.org/10.1177/1094428104268027](https://doi.org/10.1177/1094428104268027)

[https://doi.org/10.1037/a0027934](https://doi.org/10.1037/a0027934)


[https://doi.org/10.1016/j.psychres.2016.10.034](https://doi.org/10.1016/j.psychres.2016.10.034)

[https://doi.org/10.1007/bf02294825](https://doi.org/10.1007/bf02294825)

[https://doi.org/10.1007/bf02294510](https://doi.org/10.1007/bf02294510)
https://doi.org/10.1097/01.mlr.0000245438.73837.89

https://doi.org/10.1016/j.healthplace.2014.02.004

https://doi.org/10.4324/9780203821961

https://doi.org/10.1207/s15327906mbr3903_4

https://doi.org/10.31234/osf.io/u89rb


https://doi.org/10.1007/s10964-020-01295-x

measurement invariance of the SDQ across five European nations. European child & adolescent psychiatry, 24(12), 1523-1534.  
https://doi.org/10.1007/s00787-015-0729-x

https://doi.org/10.1037/1082-989x.5.3.343

https://doi.org/10.1016/j.dcn.2011.11.008

https://doi.org/10.1037/1040-3590.19.2.189

https://doi.org/10.1177/1094428103259554

https://doi.org/10.1080/00273171.2014.933762

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

https://doi.org/10.1037/0021-9010.87.3.517

https://doi.org/10.1177/2167702615611073


SDQ: Information for researchers and professionals about the strengths & difficulties questionnaires. n.d., [https://www.sdqinfo.com/](https://www.sdqinfo.com/).


