RECENT ADVANCEMENTS IN STRUCTURAL EQUATION MODELING (SEM): FROM BOTH METHODOLOGICAL AND APPLICATION PERSPECTIVES

EDITED BY: Oi-Man Kwok, Mike W.-L. Cheung, Suzanne Jak, Ehri Ryu and
Jerry Jiun-Yu Wu

PUBLISHED IN: Frontiers in Psychology

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88945-743-4
DOI 10.3389/978-2-88945-743-4

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: researchtopics@frontiersin.org

RECENT ADVANCEMENTS IN STRUCTURAL EQUATION MODELING (SEM): FROM BOTH METHODOLOGICAL AND APPLICATION PERSPECTIVES

Topic Editors:
Oi-Man Kwok, Texas A&M University, United States
Mike W.-L. Cheung, National University of Singapore, Singapore
Suzanne Jak, University of Amsterdam, Netherlands
Ehri Ryu, Boston College, United States
Jerry Jiun-Yu Wu, National Chiao-Tung University, Taiwan

Structural equation modeling (SEM) is becoming the central and one of the most popular analytical tools in the social sciences. Many classical and modern statistical techniques such as regression analysis, path analysis, confirmatory factor analysis, and models with both measurement and structural components have been shown to fall under the umbrella of SEM. Thus, the flexibility of SEM makes it applicable to many research designs, including experimental and non-experimental data, cross-sectional and longitudinal data, and multiple-group and multilevel data.

In this eBook, you will find 19 cutting-edge papers from the Research Topic: Recent Advancements in Structural Equation Modeling (SEM). These 19 papers cover a wide variety of topics related to SEM, including: (a) analysis of different types of data (from cross-sectional data with floor effects to complex survey data and longitudinal data); (b) measurement-related issues (from the development of new scale to the evaluation of person fit and new ways to test measurement invariance); and (c) technical advancement and software development. We hope that the readers will gain new perspectives and be able to apply some of the new techniques and models discussed in these 19 papers.

Citation: Kwok, O.-M., Cheung, M. W.-L., Jak, S., Ryu, E., Wu, J. J.-Y., eds. (2019). Recent Advancements in Structural Equation Modeling (SEM): From Both Methodological and Application Perspectives. Lausanne: Frontiers Media. doi: 10.3389/978-2-88945-743-4

Table of Contents

SECTION 1
ANALYSIS OF DIFFERENT TYPES OF DATA

SECTION 2
MEASUREMENT-RELATED ISSUES

SECTION 3
TECHNICAL ADVANCEMENT AND SOFTWARE DEVELOPMENT

Editorial: Recent Advancements in Structural Equation Modeling (SEM): From Both Methodological and Application Perspectives

Oi-Man Kwok[1]*, *Mike W. L. Cheung*[2], *Suzanne Jak*[3], *Ehri Ryu*[4] *and Jiun-Yu Wu*[5]

[1] Texas A&M University, College Station, TX, United States, [2] National University of Singapore, Singapore, Singapore, [3] University of Amsterdam, Amsterdam, Netherlands, [4] Boston College, Chestnut Hill, MA, United States, [5] National Chiao Tung University, Hsinchu, Taiwan

Keywords: structural equating modeling, measurement invariance (MI), multilevel (ML), Mediation, meta analysis, effect size, complex data

Editorial on the Research Topic

Recent Advancements in Structural Equation Modeling (SEM): From Both Methodological and Application Perspectives

Structural equation modeling (SEM) is becoming the central and most popular analytical tool in the social sciences. Many classical and modern statistical techniques such as regression analysis, path analysis, confirmatory factor analysis, and models with both measurement and structural components have been shown to fall under the umbrella of SEM. Thus, the flexibility of SEM makes it applicable to many research designs, including experimental and non-experimental data, cross-sectional and longitudinal data, and multiple-group and multilevel data. Further enhancing the popularity and widespread use of SEM, it has recently experienced exciting advancements—from fundamental issues like alternative estimation methods that are robust to often violated assumptions to the expansion of SEM to incorporate multilevel and cross-classified data that are common in the social sciences. This Special Research Topic aims to bring in a collection of SEM papers that not only tackle technical estimation issues but also examine and demonstrate application of SEM to more complex settings, such as applying robust estimation method, testing interaction effect, examining measurement invariance, and specifying and evaluating models applied to different types of data, including meta-analytic data, multilevel, and longitudinal data.

We are presenting 19 cutting-edge papers covering a wide variety of topics related to SEM. The papers have been grouped into three main themes: (a) analysis of different types of data (from cross-sectional data with floor effects to complex survey data and longitudinal data); (b) measurement-related issues (from the development of new scale to the evaluation of person fit and new ways to test measurement invariance); and (c) technical advancement and software development. Below you will find a summary of the three themes and corresponding papers for each.

ANALYSIS OF DIFFERENT TYPES OF DATA

One of the major advantages of SEM is its flexibility for analyzing different types of data. On this research topic, Zhu and Gonzalez have demonstrated how to analyze multilevel data with strong floor effects using multilevel SEM and examined the impact of ignoring these floor effects when using regular multilevel analysis via a Monte Carlo study. Similarly, Wu et al. have demonstrated

OPEN ACCESS

Edited and reviewed by:
Pietro Cipresso,
Istituto Auxologico Italiano (IRCCS), Italy

****Correspondence:***
Oi-Man Kwok
omkwok@tamu.edu

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *03 September 2018*
Accepted: *19 September 2018*
Published: *09 October 2018*

Citation:
Kwok O-M, Cheung MWL, Jak S, Ryu E and Wu J-Y (2018) Editorial: Recent Advancements in Structural Equation Modeling (SEM): From Both Methodological and Application Perspectives. Front. Psychol. 9:1936. doi: 10.3389/fpsyg.2018.01936

the multilevel confirmatory factor analysis with the use of complex survey data and compared different approaches to analyzing this type of data. Their simulation results showed that the maximum modeling strategy generally outperforms the other approaches.

In addition, several papers focus on the analysis of longitudinal data from different perspectives. Ning and Luo have introduced and evaluated a new piecewise growth-curve model (PGCM) without the requirement of pre-specifying the turning point. Similarly, Kim et al. have proposed an optimal starting model under the latent growth modeling (LGM) framework when searching for the accurate growth trajectory. These authors found that the fully saturated model performed the best even with the presence of the time-invariant covariates in the LGM (i.e., the conditional LGM). Kamata et al. have investigated the performance of three approaches (i.e., one-step, three-step, and case-weight) on estimating a two-phase mixture model with an auxiliary linear-growth model. This simulation study showed that under different conditions some approaches outperformed the others (e.g., both case-weight and three-step resulted in higher convergence rate but could also lead to substantially underestimated standard errors when the class separation was low). As an extension of the mixture model, the growth-mixture model (GMM) is another commonly used model for analyzing longitudinal data. Focusing on GMM, Kim and Wang have conducted two Monte Carlo studies to examine the impact of ignoring the presence of measurement non-invariance between latent classes in terms of class enumeration and parameter recovery when applying both GMM and second-order GMM. In general, the second-order GMM outperformed the traditional GMM with more accurate class enumeration and unbiased parameter estimates. For more complex longitudinal data such as students moving to different classrooms over time, Kwok et al. have demonstrated how to analyze this type of data, especially in terms of capturing the carry-over effect, with the use of the Project ELLA data along with the xxM program.

MEASUREMENT-RELATED ISSUES

Measurement models are an important part of SEM, and the flexibility of SEM not only allows researchers to develop and validate new scales but also provides a simple and feasible platform for examining the potential differences between groups and populations through the test of measurement invariance. Zhao et al. have developed and validated their online shopping addiction scale with both exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). Similarly, Glaman and Chen have tested the measurement invariance of a classroom engagement measure among academically at-risk students across grades, genders, and ethnicities. In addition, several excellent papers address the methodological issues involved in testing measurement invariances, including Jorgensen's work on using permutation tests and multivariate modification indices, Jiang et al.'s study on using equivalence tests with a projection-based approach on testing measurement invariance and mean comparison, and Hsiao and Lai 's paper on examining the impact of partial measurement invariance on testing moderation for single and multilevel data.

By extending the measurement test to more complex data settings, Jak and Jorgensen have showed the relationship between measurement invariance, cross-level invariance, and multilevel reliability. Moreover, Guenole has evaluated the impact of biased-referent indicators with the use of free vs. constrained baseline approaches within a multilevel SEM framework. His simulation results re-emphasize the importance of having an unbiased referent indicator when testing measurement invariance.

TECHNICAL ADVANCEMENT AND SOFTWARE DEVELOPMENT

This special research topic includes several articles on new technical advancement and software development in SEM. For example, effect size reporting becomes necessary and is required by most of the prestigious peer-reviewed journals in behavioral and social sciences. Cheung has addressed the importance of the multivariate effect sizes and demonstrated how to compute multivariate effect sizes and the corresponding covariance matrices under the SEM framework with the use of the metaSEM package.

The normal-theory maximum likelihood (ML-Normal) is the most commonly used estimation method in SEM and is also the default estimation method for most SEM-related software. However, ML-Normal is not efficient and can be severely biased by outliers and influential observations in the data. Lai and Zhang have evaluated the performance of the fit indices from the multivariate *t*-based SEM framework and recommend that the multivariate *t*-based SEM be used when outliers and influential observations exist in the data. Similarly, given that linear factor analysis (FA) is another commonly used SEM approach for psychometric applications, Ferrando et al. have proposed a simple and workable approach that can routinely assess person fit in FA-based studies. Through both simulation study and real-data demonstration, they found that the mean-squared *lico* index and the personal correlation work well in conjunction and can function effectively for detecting different types of inconsistency.

Mediation or indirect effect is fundamental to many substantive areas. For comparing indirect effects in different groups, Ryu and Cheong have examined both single-group and multiple-group SEM approaches, concluding that the multiple-group approach is generally the preferable approach. They also recommend the use of the bootstrap confidence intervals when adopting the single-group approach. In a similar vein, confirmatory factor analysis (CFA) is commonly used in the social sciences, but specifying the model is sometimes tricky, especially for complex data settings such as multilevel data. Hence,Wu et al. have developed an integrated MCFA (iMCFA) program that allows researchers to easily and flexibly fit the single-level CFA models as well as the multilevel CFA models with maximum model at either the within- or the between-level.

We hope that the readers will gain new perspectives and be able to apply some of the new techniques and models discussed in the 19 papers in this special research topic. Moreover, we hope that more advanced readers will conduct more exciting studies and move the field by extending some of the methods and ideas from the papers in this special topic.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 12 December 2017
doi: 10.3389/fpsyg.2017.02146

Modeling Floor Effects in Standardized Vocabulary Test Scores in a Sample of Low SES Hispanic Preschool Children under the Multilevel Structural Equation Modeling Framework

Leina Zhu and Jorge Gonzalez*

Psychological, Health & Learning Sciences, University of Houston, Houston, TX, United States

OPEN ACCESS

Edited by:
Jiun-Yu Wu,
National Chiao Tung University, Taiwan

Reviewed by:
Okan Bulut,
University of Alberta, Canada
Xing Liu,
Eastern Connecticut State University, United States

****Correspondence:***
Leina Zhu
zhu_leina@163.com

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *10 May 2017*
Accepted: *27 November 2017*
Published: *12 December 2017*

Citation:
Zhu L and Gonzalez J (2017) Modeling Floor Effects in Standardized Vocabulary Test Scores in a Sample of Low SES Hispanic Preschool Children under the Multilevel Structural Equation Modeling Framework. Front. Psychol. 8:2146. doi: 10.3389/fpsyg.2017.02146

Researchers and practitioners often use standardized vocabulary tests such as the Peabody Picture Vocabulary Test-4 (PPVT-4; Dunn and Dunn, 2007) and its companion, the Expressive Vocabulary Test-2 (EVT-2; Williams, 2007), to assess English vocabulary skills as an indicator of children's school readiness. Despite their psychometric excellence in the norm sample, issues arise when standardized vocabulary tests are used to asses children from culturally, linguistically and ethnically diverse backgrounds (e.g., Spanish-speaking English language learners) or delayed in some manner. One of the biggest challenges is establishing the appropriateness of these measures with non-English or non-standard English speaking children as often they score one to two standard deviations below expected levels (e.g., Lonigan et al., 2013). This study re-examines the issues in analyzing the PPVT-4 and EVT-2 scores in a sample of 4-to-5-year-old low SES Hispanic preschool children who were part of a larger randomized clinical trial on the effects of a supplemental English shared-reading vocabulary curriculum (Pollard-Durodola et al., 2016). It was found that data exhibited strong floor effects and the presence of floor effects made it difficult to differentiate the invention group and the control group on their vocabulary growth in the intervention. A simulation study is then presented under the multilevel structural equation modeling (MSEM) framework and results revealed that in regular multilevel data analysis, ignoring floor effects in the outcome variables led to biased results in parameter estimates, standard error estimates, and significance tests. Our findings suggest caution in analyzing and interpreting scores of ethnically and culturally diverse children on standardized vocabulary tests (e.g., floor effects). It is recommended appropriate analytical methods that take into account floor effects in outcome variables should be considered.

Keywords: ethnically and culturally diverse children, standardized vocabulary tests (the PPVT-4, the EVT-2), floor effects, intervention effects

INTRODUCTION

The Peabody Picture Vocabulary Test–IV (PPVT-4; Dunn and Dunn, 2007) and the Expressive Vocabulary Test–II (EVT-2; Williams, 2007), along with their earlier versions, are the most widely used standardized vocabulary tests in the United States and other countries, including: Slovania (e.g., Bucik and Bucik, 2003), France (e.g., Theriault-Whalen and Dunn, 1993), Japan (e.g., Ueno et al., 1991), Korea (e.g., Kim et al., 1995), Brazil (e.g., Capovilla and Capovilla, 1997), Northern Sotho (e.g., Pakendorf and Alant, 1997), and China (e.g., Ji et al., 2014). The popularity of these measures is evidenced by over 1,000 combined citations from 1960 to 2016 in PSYCHINFO alone. Nevertheless, debates and criticisms over use of these vocabulary tests with culturally and linguistically diverse populations continue unabated.

Criticisms of standardized vocabulary tests have ranged from content bias, bias in reference norms, threats to content and construct validity, to cultural bias and so forth (e.g., Stockman, 2000; Qi et al., 2003; Thomas-Tate et al., 2006; Haitana et al., 2010; Pae et al., 2012). Among the most salient criticism is the use of these tests with children from low-income, culturally, ethnically, and linguistically diverse backgrounds. Among vocabulary measures, The PPVT is among the most popular. The PPVT is a standardized measure of children's receptive vocabulary and screen for verbal abilities. The use of the PPVT is widespread including use in large scale federal funded early childhood programs including Even Start Programs and Early Reading First and use by speech-language pathologists for verbal ability evaluations. Its companion the EVT measures children's expressive vocabulary and complements the PPVT (Restrepo et al., 2006). The PPVT, in particular, has sparked much controversy over alleged inappropriateness with culturally and linguistically diverse populations (Haitana et al., 2010). Numerous studies have shown ethnically and linguistically diverse populations to score one to two standard deviations below normative expectations (e.g., Washington and Craig, 1992, 1999; Champion et al., 2003; Laing and Kamhi, 2003; Qi et al., 2003; Restrepo et al., 2006; McCabe and Champion, 2010; Terry et al., 2013; Gonzalez et al., 2015), highlighting possible bias in these tests. African-American, Hispanic and Native American populations, in particular, have been shown to score much lower on standardized vocabulary tests than do the normative samples (Thernstrom, 2002; Buly, 2005; Rock and Stenner, 2005; Thomas-Tate et al., 2006; Horton-Ikard and Ellis Weismer, 2007). African American children, for example, have been shown to score about one standard deviation below the mean scores compared to their White counterparts (e.g., Rock and Stenner, 2005; Restrepo et al., 2006). Latino preschoolers have been found to approach two or more standard deviations below normative standards (Lonigan et al., 2013; Gonzalez et al., 2015).

The suitability of the PPVT-4 and EVT-2 for use with ethnically, linguistically or culturally different populations continues debated. As highlighted in the manual, the PPVT-4 and EVT-2 were developed to measure standard American English (Dunn and Dunn, 2007; Williams, 2007)-a potential bias for non-standard English speaking or English learning populations. For example, neither the PPVT nor the EVT incorporate African American English (dialect of American English) in the test items (Qi et al., 2003; Pae et al., 2012). Researchers have also questioned the use of a predominately White middle-class American norm sample in both tests (Qi et al., 2003). The predominantly White norms of both tests have raised concerns in their use when testing cultural and ethnical diverse groups (Stockman, 2000; Thomas-Tate et al., 2006; Haitana et al., 2010).

Examining the appropriateness of standardized vocabulary tests for use with linguistically, culturally or ethnically different populations remains a high priority. Additionally, few efforts have been made to address how researchers can analyze data from non-English or non-standard English speaking children in ways that take into account possible biases in the tests. As discussed previously, culturally or ethnically different populations generally score disproportionately lower than the normative sample on standardized vocabulary tests. Many among these populations score at the lower end of the distribution of scores. In psychology and social science research, when test scores "stack" on or near the lower end of measurement scale, this phenomenon is known as "floor effects" (Hessling et al., 2004; McBee, 2010). Notable among tests that yield floor effects among ethnically and linguistically diverse populations are the PPVT-4 or the EVT-2. Researchers or others using the PPVT or EVT tests need to be aware that relative to the norming sample, scores for culturally and linguistically different populations may show right-skewed data distribution patterns or floor effects. Among the concerns with skewed data patterns is that many parametric statistical analytic strategies (e.g., *t*-test, ANOVA, and multiple regression) rely on normality assumptions. Inappropriate data analytical strategies result in distorted results and quite possibly erroneous or incorrect inferences due to violations of model assumption. Given that highly skewed distributions of floor effects, the use of conventional statistical methods assuming normality may yield distorted and quite possibly misleading results (Muthén and Asparouhov, 2010). For example, Hessling et al. (2004) point out that due to floor effects experimental and quasi-experimental intervention studies may fail to reject the null results when in fact the null hypothesis is rejected. As an example, if there is insufficient range in the measurement scale to capture and differentiate lower levels of ability or achievement, low-performing participants will tend to score in or "stack" at the low end of the scale. In such situations, the presence of floor effects renders it difficult to compare the invention group with the control group in terms of gains produced by an intervention. In sum, floor effects may distort efforts at examination of intervention effects, in particular, among diverse populations such as children from low-income, culturally, ethnically, and linguistically diverse background.

The importance of addressing floor effects in data analyses is largely undisputed. In both simulation and empirical studies researchers have demonstrated that ignoring floor effects can result in biases in parameter estimates, standard error estimates, and misleading inferences (Wang et al., 2008; Twisk and Rijmen, 2009; McBee, 2010). To address potential floor effects in data analysis, techniques to deal with the floor and similar type of data have been developed and increasingly applied in social science research (e.g.,Twisk and Rijmen, 2009; McBee, 2010; Proust-Lima et al., 2011; Iachina and Iachina, 2012; Whitaker and Gordon, 2012; Keeley et al., 2013). For example, the practice of treating floor data as left-censored data and using the Tobit regression model as a correction has been a common recommendation (e.g., Cox and Oakes, 1984; Muthén, 1989, 1990; Klein and Moeschberger, 1997). The concept of floor effects is similar to left-censoring in survival analysis framework. In survival analysis, left censoring is considered to when some individuals have already experienced the event of interest before recording or observing or collecting those targeted data points (Kleinbaum and Klein, 2005). Floor effects are in similar nature. Due to a measurement range that does not adequately capture extremely low levels of ability and/or achievement, some true scores beyond the scale limits cannot be observed, similar to the left-censored data which are censored/truncated at the lower-boundary (floor threshold). While left censoring is related to the observation time, floor effects are in the context of restricted range of measurement. In Tobit regression model (also called censored regression), from treating floor data as left-censored data, Tobit regression effectively models the limitation.

Recognizing that floor effects in data analysis can lead to biased estimates, it is recommended that researchers more closely examine the distribution of scores in standardized measures administrated to diverse populations. For example, in a sample of students with special needs, Whitaker (2005, 2008, 2010, 2012) identified possible floor effects in their scores on both the Wechsler Adult Intelligence Scale (WAIS) and the Wechsler Intelligence Scale for Children (WISC). Similarly, while screening a large cohort of students for reading disabilities, Catts et al. (2009) pointed out floor effects in their scores on the Dynamic Indicators of Basic Early Literacy Skills (DIBELS), a screening instrument for identifying children at risk for reading disabilities. Many children were found to score near the lower end of the distribution (no or low risk for reading disabilities). These studies demonstrate that floor effects may occur when these measures are used with diverse groups. Nevertheless to our knowledge, no studies exist examining the impact of floor effects in administrations of the PPVT-4 or EVT-2 with children from low-income, culturally, ethnically, and linguistically diverse backgrounds. As discussed previously, numerous studies found these children often performed poorly on the PPVT-4 or EVT-2 with the vast majority of scores stacked near the lower end of the data distribution. While there is no universally accepted definition of what constitutes floor effects in tests, in some disciplines (e.g., clinical orthopedics research), floor effects are defined as when 15% (or more) of sample participants score at the lowest level of a measure's range (Lim et al., 2015). Given the predominance of low scores for cultural and ethnical diverse groups on the PPVT-4 or EVT-2, particular attention needs to be paid to the presence of floor effects in the data.

THE PURPOSE OF THE STUDY

Despite psychometric excellence of the PPVT-4 and the EVT-2, concerns arise when these tests are used to asses diverse populations who may perform substantively different from the norm sample. As noted in research, non-English or non-standard English speaking children often score one to two standard deviations below normative standards (e.g., Lonigan et al., 2013). This study focused on Mexican-American Spanish-speaking preschool dual language learners (DLL) enrolled in preschool.

The study had three aims: (a) to examine floor effects in data from a the pre-test administration of the PPVT-4 (Dunn and Dunn, 2007) and the EVT-2 (Williams, 2007) test scores in a sample of low SES Mexican-American DLL preschool children (Pollard-Durodola et al., 2016), (b) to examine the impact of floor effects on evaluating the pre- post-test performance on receptive and expressive vocabulary outcomes as measured by the PPVT-4 and the EVT-2 (Pollard-Durodola et al., 2016), and (c) to evaluate the impact of floor effects on estimating parameters, standard errors, and significant tests through Monte Carlo simulations. Different analytical approaches were compared in response to different levels of floor data in the outcome variable in the multilevel structural equation modeling (MSEM) framework, which is viewed as a more general framework to analyze multilevel data. Results discussed and appropriate statistical methods for dealing with data with floor effects were thereby suggested.

DEALING WITH FLOOR EFFECTS

In this study, we examined three methods of analyzing data from pre and post administration of the PPVT-4 and EVT-2 with potential floor effects, including the regular multilevel regression model with maximum likelihood (ML) estimation (ignoring floor effects), the robust standard error approach in multilevel model (standard error adjustment based on maximum likelihood with robust standard error estimation), and the multilevel Tobit regression model (addressing floor effects from treating the outcome variable with floor effects as left-censored variable). All these analyses are set up under the multilevel structural equation modeling (MSEM) framework given that MSEM is viewed as a more general framework for analyzing multilevel data with the flexibility to include both observed and latent variables in the model simultaneously (Muthén and Muthén, 1998-2015). Conventional linear regression assumes normality assumption. Floor effects in the dependent variable are not taken into account in the conventional linear regression analysis (Winship and Mare, 1984). The robust standard error approach and the Tobit approach, on the other hand, handle floor effects with different techniques. In the next section, the latter two approaches are presented in more detail.

The Robust Standard Error Approach

Statistical methods often rely on certain assumptions, such as multivariate normality, homoscedasticity, or observation independency. If model assumptions are not satisfied, substantial biases would occur in parameter estimates, standard error estimates, and model evaluation. Floor effects generally occur when data distributions are highly right skewed. Given floor effects in the outcome variables, the use of linear regression is problematic due to potential violation of the multivariate normality assumption. Yuan et al. (2005), for example, demonstrated that standard error estimates and test statistics may be inconsistent due to data nonnormality (e.g., positive skewed data). Brown (2006) noted marked floor effects led to biased standard error estimates using maximum likelihood (ML). Note that in some conditions, normal theory ML produced unbiased parameter estimates though data are nonnormal, however, bias in standard error estimates cannot be overcome and possibly distorting significance testing, and in turn misleading inferences (Yuan and Bentler, 2000; Finney and DiStefano, 2006; Baraldi and Enders, 2010).

To correct for bias in standard error estimates, robust standard error approach has often been used to produce unbiased standard errors (King and Roberts, 2015). The literature identifies several ways to obtain robust standard errors, such as asymptotically distribution-free estimation (ADF; Browne, 1984) and bootstrapping (Nevitt and Hancock, 2001). Other methods include Huber/Pseudo sandwich estimator. In the M*plus* program (Muthén and Muthén, 1998-2015), there are three routines to produce "robust" standard errors, including: (1) maximum likelihood parameter estimates with robust standard errors and chi-square test statistic (MLM), (2) maximum likelihood parameter estimates with standard errors and a mean- and variance-adjusted chi-square test statistic that are robust to non-normality (MLMV), and (3) maximum likelihood parameter estimates with standard errors and chi-square test statistic robust to non-normality and observation non-independence (MLR). In this study, M*plus* was used for all analyses and illustrations.

The three estimation methods, namely, MLM, MLMV, and MLR, are all ML based robust estimators. However, standard errors produced by these ML estimators could be very divergent. In many situations, the ML parameter estimates are still consistent even data are nonnormal, but standard error estimates could be very biased. In analyzing multilevel data, MLR shows its advantage in dealing with observation dependency (Maas and Hox, 2004). In addition, MLR is also superior in handling: (1) data non-normality and (2) missing data (see Yuan and Bentler, 2000). In this study, we adopted MLR estimator in terms of handling both floor effects in data and data of multilevel structure.

The Tobit Approach

Tobit regression analysis, first formulated by Tobin (1958), models linear relationships between variables when the outcome variable is either a left- or right-censored variable. In Tobit regression, scores that fall at or below some threshold are viewed to be (left) censored from below the threshold. As described previously, floor effects are potential when a large percentage of scores occurs at the low end of the measurement scale. Data with floor effects are treated as left-censored data in Tobit regression. For instance, when two low-performing students are measured with a standardized test, both students scored zero on the test, but their actual abilities may not be the same. In this case, their scores seem to be censored from the censoring point (i.e., zero), which however, fail to capture their true abilities. The standardized test, because of its restricted score range, is unable to differentiate abilities of students who score extremely low (or high) level. Scores at the extremes can be viewed as being censored or truncated. The lowest (or highest) bound is called the censoring point or threshold (Cox and Oakes, 1984). To sum, in the Tobit regression model dependent variables with floor effects are viewed as left-censored variables.

In the Tobit regression model, y^* represents a random latent variable and y represents a censored variable. When the data are not censored, the distributions of y^* and y overlap. The lowest bound is defined as "l" and the highest bound as "u". Mathematically, the Tobit regression models are expressed as follows: (Long, 1997; Twisk and Rijmen, 2009):

$$y_i^* = \beta_0 + \beta_1 x_i{}' + e_i, e_i \sim N(0, \sigma^2), \quad (1)$$

$$y_i = l \text{ for } y_i^* \leq l, \quad (2)$$

$$y_i = y_i^* \text{ for } l < y_i^* \quad (3)$$

when the outcome variable is left-censored.

When the outcome variable is right-censored, expressions include

$$y_i^* = \beta_0 + \beta_1 x_i{}' + e_i, e_i \sim N(0, \sigma^2), \quad (4)$$

$$y_i = y_i^* \text{ for } y_i^* < u, \quad (5)$$

$$y_i = u \text{ for } y_i^* \geq u. \quad (6)$$

The discussion till now was about a simple Tobit regression model. When data are characterized by dependency among observations due to the nested or hierarchical data structure (e.g., students nested within classrooms, members nested within organizations), multilevel model (MLM) is the appropriate method (Raudenbush and Bryk, 2002; Hox et al., 2010). The following expressions (7–9) represent a typical multilevel model (i.e., random intercept model which is equivalent to a commonly used form of multilevel structural equation model (MSEM) and can be specified and analyzed by the M*plus* Type = Twolevel routine):

$$\text{Level 1}: \mathrm{Y}_{ij} = \beta_{0j} + \beta_{1\mathrm{j}}\mathrm{X}_{ij} + \mathrm{e}_{ij}, \quad (7)$$

$$\text{Level 2}: \beta_{0\mathrm{j}} = \gamma_{00} + \mathrm{U}_{0j}, \quad (8)$$

$$\beta_{1j} = \gamma_{10}, \quad (9)$$

where i represents the individual (i.e., $i = 1\ldots n_{\mathrm{j}}$) and j represents the group in which the individual is nested (i.e., $j = 1\ldots N$). In the level-1 model as shown in Equation (7), β_{0j} is the estimated average for the j-th group. B_{1j} is the slope which is a fixed effect and X_{ij} is the level 1 covariate. e_{ij} is the within-group random error. In the level-2 models shown in Equations (8) and (9), β_{0j} is the random intercept constituted by the grand mean (γ_{00}) and the between-group random effect (U_{0j}). We interpret the estimate for U_{0j} as the variance of the mean for each group around the grand mean. In Equation (9), given the slope is a fixed effect, γ_{10} represents the average change across all groups for the X_{ij} predictor.

To address floor effects in the outcome variable, the above Equations (7), (8), and (9) can be modified and the following equations represent a multilevel Tobit regression model:

$$\text{Level 1}: y_{ij}^{*} = \beta_{0j} + \beta_{1j}x_{ij}{}' + e_{ij}, \quad (10)$$

$$y_{ij} = l \text{ for } y_{ij}^{*} \leq l, \quad (11)$$

$$y_{ij} = y_{ij}^{*} \text{ for } l < y_{i}^{*} \quad (12)$$

$$\text{Level 2}: \beta_{0j} = \gamma_{00} + U_{0j}, \quad (13)$$

$$\beta_{1j} = \gamma_{10}, \quad (14)$$

in which the outcome variable with floor effects (y_{ij}) is treated as left-censored (y_{ij}^{*}) in the multilevel Tobit regression model.

Next, three comparative methods, including regular multilevel regression, multilevel regression with robust standard error approach, and multilevel Tobit regression, were examined in response to floor effects in data with multilevel structure. The first method did not address the floor effects. The latter two methods addressed the floor effects differently. For the robust standard error approach, the MLR estimator was adopted to obtain robust standard errors. For the multilevel Tobit regression approach, the outcome variable was treated as a left-censored variable in which a multilevel Tobit regression was applied. Next, we presented two studies to examine the three methods. First an empirical example was presented, followed by a simulation study.

AN EMPIRICAL EXAMPLE

Although there is no consensus, standardized vocabulary tests (e.g., the PPVT-4 and the EVT-2) may, under some circumstances, be inappropriate for use with culturally or linguistically diverse populations. Given evidence of low standardized vocabulary tests scores from some cultural and linguistically diverse groups, this study highlighted the potential issue of floor effects. In summary, the aim of this empirical example was 2-fold: (a) to establish the existence of floor effects on the PPVT-4 and the EVT-2 test scores in a sample of low SES Mexican-American preschool children who were dual language learners (DLLs) (Pollard-Durodola et al., 2016), and (b) to investigate the impact of floor effects on examining the PPVT-4 and the EVT-2 pre- to post-test scores comparison with respect to the sample's vocabulary growth (Pollard-Durodola et al., 2016).

In this example, the participants included 252 low-income Mexican-American preschool children participating in randomized clinical trial of an evidence-based shared book reading intervention in two school districts located in South Texas. In this sample, preschool children (average age was 5 years) were 92.1% economically disadvantaged, and primarily Mexican-American (98.3%). Eighty-seven percent of parents of preschoolers reported that Spanish was the primary language spoken at home while 8% reported speaking English in the home and 5% reported using both languages. All children were identified as Spanish-speaking children while learning English as a second language (Pollard-Durodola et al., 2016). Based on the student performance on the *pre*LAS® English (DeAvila and Duncan, 2000), all preschoolers were at the pre-functional and beginning level for their English language proficiency. All the preschool children were assessed using the PPVT-4 and the EVT-2 at pre- and posttests to examine the impact of the shared-reading intervention on their vocabulary growth.

Table 1 provides descriptive on standardized scores for the post-test PPVT-4 and the EVT-2. As shown in **Table 1**, a significant majority of participants scored in the low range, including: 93.97% on the PPVT-4 (i.e., moderately low range 33.73% + extremely low range 60.24%) and 92.01% on the EVT-2 (i.e., moderately low range 31.09% + extremely low range 60.92%). The mean score on PPVT-4 was 63.81 (*SD* = 15.42), corresponding to two standard deviations (*SD*s) below the normative mean of 100. The mean score on the EVT-2 was 55.63 (*SD* = 24.01), corresponding to three standard deviations (*SD*s) below the normative mean of 100 The standardized scores on the PPVT-4 ranged from 20 to 91, which indicated that all participants (*N* = 252) scored below the normative mean (i.e., 100). On the EVT-2, the scores ranged from 20 to 108. Only two out of 252 participants (i.e., 1%) scored above the normative mean (i.e., 100) while 99% (i.e., *n* = 250 out of 252) scored below the normative mean (i.e., 100). In summary, post-test standardized scores on the PPVT-4 and EVT-2 for the sample of low SES Mexican-American preschool children suggested evidence of floor effects.

Figure 1 displays the distributions of standardized PPVT-4 and EVT-2 scores of the sample of preschool children, respectively. The distributions demonstrate that a preponderance of scores fell in the low ranges, especially on the EVT-2. As noted earlier, if 15% or more of the sample scores in the lowest level of a measure range, floor effects likely exist. Regardless of the standardized scores or the raw scores, this sample of Mexican-American preschool children performed significantly lower relative to the norm sample on the PPVT-4 and the EVT-2 with the vast majority scoring on or near the low end of measurement scale.

In the second aim of this study, we explored the influence of floor effects on the pre- to post effectiveness of the shared-reading intervention on the sample of Mexican-American children. Specifically, we wanted to know whether floor effects masked vocabulary growth. As shown in **Table 2**, at pre-test the children scored on average two standard deviations (*SD*s) below the normative mean on the PPVT-4 and three standard deviations (*SD*s) below the normative mean on the EVT-2. The children

TABLE 1 | Distribution of scores and corresponding descriptive for all of the participants.

Descriptor	Standard score range	Measures			
		PPVT-4 (*N* = 249)		EVT-2 (*N* = 238)	
		n	Percentage	*n*	Percentage
Extremely high	130+	0	0	0	0
Moderately high	115–129	0	0	0	0
High average	100–114	0	0	2	1
Low average	85–99	15	6	17	7
Moderately low	70–84	84	34	74	31
Extremely low	≤69	150	60	145	61

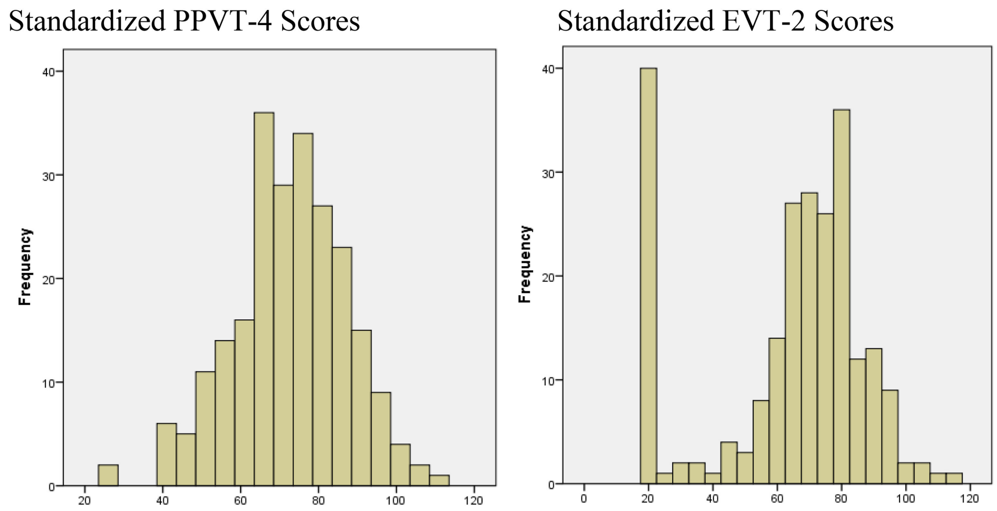


FIGURE 1 | Distribution of standardized vocabulary test scores of a sample of low SES Hispanic preschool children.

TABLE 2 | Pretest and posttest scores for intervention and comparison groups.

Measure	Pretest				Posttest			
	Total	Intervention	Comparison	*t*	Total	Intervention	Comparison	*t*
PPVT-4								
N	249	136	113	0.05, $p = 0.957$	234	129	105	0.53, $p = 0.599$
M	63.81	63.86	63.75		72.70	73.16	72.13	
SD	15.42	14.76	16.24		14.63	14.12	15.29	
EVT-2								
N	238	132	106	0.17, $p = 0.864$	232	127	105	−0.30, $p = 0.762$
M	55.63	55.87	55.33		64.08	63.65	64.60	
SD	24.01	23.59	24.64		24.01	25.09	22.75	

PPVT-4, Peabody Picture Vocabulary Test (4th ed.); EVT-2, Expressive Vocabulary Test (2nd ed.).

still lagged behind at posttests on average showing one standard deviation (*SD*s) below the normative mean on the PPVT-4 and two standard deviations (*SD*s) below the normative mean on the EVT-2. When having the first glance at the post-test scores, there appeared to be no difference between the intervention and control groups on the two standardized vocabulary measures: PPVT-4: $t = 0.53$, $p = 0.599$; EVT-2: $t = -0.30$, $p = 0.762$. According to the pretest-posttest comparison, one could reasonably conclude that the intervention had been ineffective in accelerating vocabulary growth for the treatment group of children. One interpretation would be that the shared book reading intervention designed to show promise in improving children's vocabulary for diverse children was not effective. Results must, however, be interpreted in light of the staggeringly poor performance of the Mexican-American preschoolers at pretest (e.g., many children scored two to three standard deviations below monolingual vocabulary norms). Because their pre-test vocabulary performance was so low, it appears that these preschool children were unresponsive to the intervention. In this scenario analyzing performance using conventional analytic methods on the preschool children's pre- to post-test PPVT-4 and EVT-2 without adequately taking into account of the floor effects may have resulted in misleading conclusions about the effectiveness of the intervention.

Table 3 presents model results using the three different methods (i.e., the traditional multilevel model without addressing the floor effects, the robust standard error approach which partially addressing the floor effects, and the multilevel Tobit model which directly addressing the floor effects). An annotated input from the M*plus* program for analyzing a multilevel Tobit model was presented in Appendix A. As shown in **Table 3**, the results showed a mixed pattern, for example, some parameter estimates appeared to be larger when floor effects were considered (e.g., intervention) whereas some other estimates tended to be smaller (e.g., pretest, *pre*LAS® English). Specifically, for the EVT-2 outcome (with stronger floor effects compared with the PPVT-4 outcome), the approaches which accounting for floor effects yielded larger parameter estimates (e.g., gender, intervention, and years of teaching). Regarding the standard error estimates, methods

TABLE 3 | Results of hierarchical linear model to the low SES Hispanic Latino preschool children with or without modeling floor effects.

Parameter estimates and standard errors	**Dependent variable**					
	PPVT-4			**EVT-2**		
	Multilevel model floors ignored	**Robust standard error**	**Multilevel model floors considered**	**Multilevel model floors ignored**	**Robust standard error**	**Multilevel model floors considered**
FIXED EFFECTS						
Intercept	35.97*	35.97*	36.93*	1.62	1.62	0.51
(*SE*)	(13.46)	(14.54)	(15.42)	(3.10)	(3.13)	(2.84)
Level-1 Pretest (γ_{10})	0.98*	0.98*	0.98*	0.83*	0.83*	0.82*
(*SE*)	(0.01)	(0.01)	(0.01)	(0.03)	(0.03)	(0.04)
Level-1 Age (γ_{20})	−0.01	−0.01	−0.01	−0.002	−0.002	−0.01
(*SE*)	(0.02)	(0.02)	(0.02)	(0.05)	(0.04)	(0.04)
Level-1 Gender[a] (γ_{30})	−0.004	−0.004	−0.004	−0.45	−0.05	−0.03
(*SE*)	(0.02)	(0.02)	(0.02)	(0.04)	(0.04)	(0.04)
Level-1 Bilingual[b] (γ_{40})	−0.04	−0.04	−0.04	−0.05	−0.05	−0.05
(*SE*)	(0.03)	(0.02)	(0.03)	(0.06)	(0.06)	(0.06)
Level-1 Ethnicity[c] (γ_{50})	0.003	0.003	0.003	−0.02	−0.02	−0.02
(*SE*)	(0.02)	(0.01)	(0.01)	(0.04)	(0.04)	(0.05)
Level-1 Attendance (γ_{60})	0.02	0.02	0.02	0.01	0.01	0.01
(*SE*)	(0.02)	(0.02)	(0.02)	(0.04)	(0.04)	(0.05)
Level-1 *pre*LAS® English (γ_{70})	0.000	0.000	0.000	0.05	0.05	0.04
(*SE*)	(0.02)	(0.02)	(0.02)	(0.05)	(0.04)	(0.05)
Level-1 *pre*LAS® Spanish (γ_{80})	0.001	0.001	0.002	−0.01	−0.01	−0.01
(*SE*)	(0.02)	(0.02)	(0.02)	(0.05)	(0.04)	(0.05)
Level-2 School district[d] (γ_{01})	−0.70	−0.70	−0.73	−0.58	−0.58*	−0.54
(*SE*)	(0.46)	(0.37)	(0.39)	(0.39)	(0.25)	(0.23)
Level-2 Intervention (γ_{02})	***0.24***	***0.24***	***0.25***	***0.40***	***0.40****	***0.41****
(SE)	***(0.25)***	***(0.22)***	***(0.24)***	***(0.21)***	***(0.19)***	***(0.19)***
Level-2 Teacher's primary language[e] (γ_{03})	−0.59*	−0.59*	−0.60*	−0.35	−0.35	−0.31
(*SE*)	(0.28)	(0.23)	(0.24)	(0.22)	(0.23)	(0.22)
Level-2 Years of teaching (γ_{04})	−0.81*	−0.81*	−0.81*	−0.44	−0.44	−0.39
(*SE*)	(0.40)	(0.40)	(0.41)	(0.32)	(0.40)	(0.40)
Level-2 Years of teaching in PreK (γ_{05})	−0.22	−0.22	−0.23	−0.52	−0.52	−0.57
(*SE*)	(0.34)	(0.41)	(0.42)	(0.29)	(0.31)	(0.33)
Level-2 University reading credits (γ_{06})	1.11*	1.11*	1.14*	0.86*	0.86*	0.87*
(*SE*)	(0.40)	(0.37)	(0.39)	(0.30)	(0.34)	(0.32)
Level-2 Professional development (γ_{07})	0.03	0.03	0.06	0.63	0.63*	0.57
(*SE*)	(0.45)	(0.39)	(0.40)	(0.37)	(0.27)	(0.26)
RANDOM EFFECTS						
Level-1 Residual Variance (σ^2)	0.06*	0.06*	0.06*	0.29*	0.29*	0.32*
(*SE*)	(0.01)	(0.01)	(0.01)	(0.04)	(0.05)	(0.05)
Level-2 Residual Variance (τ_{00})	0.23	0.23	0.19	0.46	0.46*	0.48*
(*SE*)	(0.39)	(0.34)	(0.37)	(0.25)	(0.20)	(0.20)

*PPVT-4, Peabody Picture Vocabulary Test (4th ed.); EVT-2, Expressive Vocabulary Test (2nd ed.). *The significance level is set at $p < 0.05$ (two-tailed).*

[a] *The reference group for gender is female (coded 0).*

[b] *The reference group for bilingual is non-bilingual (coded 0).*

[c] *The reference group for ethnicity is Native American (coded 0).*

[d] *The reference group for school district is school district A (coded 0).*

[e] *The reference group for teachers' primary language is English (coded 0).*

Bold and italic values indicated a contrast of significant effects to non-significant effects when floor effects were addressed regarding the intervention effects which is the target research interest in the empirical study.

addressing floor effects generally produced smaller standard errors than the traditional multilevel model which ignoring floor effects.

The most intriguing findings in **Table 3** were the potential influence of floor effects on testing the intervention effects (i.e., γ_{02} in **Table 3**). Non-significant intervention effects were detected for the PPVT-4 and EVT-2 outcomes when using the regular multilevel regression without addressing the floor effects. Nevertheless, both robust standard error approach (partially addressing the floor effects) and Tobit regression approach (fully addressing the floor effects) yielded significant intervention effects on the EVT-2, the measure of expressive vocabulary, but non-significant intervention effects on the PPVT-4, the receptive vocabulary measure. By further examining the descriptive statistics as shown in **Table 1**, we found that more children scored near the lower end of the EVT-2 than on the PPVT-4. These results validated that the necessity of taking the floor effects into account when conducting the data analysis with potential floor effects. Without properly addressing the floor effects, one can result in the incorrect test of the significant intervention effect and mislead to the non-significant intervention effect conclusion.

In summary, floor effects were shown to be present in both standardized receptive and expressive vocabulary tests scores in a sample of low income Mexican-American preschool children who enrolled in a randomized clinical trial of a shared-book reading intervention (Pollard-Durodola et al., 2016). Analytical methods ignoring the floor effects (i.e., regular multilevel regression) and methods addressing the floor effects (i.e., robust standard error approach partially addressing the floor effects and Tobit regression approach fully addressing the floor effects) resulted in difference in model results. Accounting for floor effects in data analysis yielded different results (i.e., standard error estimates and significance tests), though parameter estimates did not appear to be significantly impacted. When floor effects were ignored, standard errors tended to be overestimated. On the other hand, both robust standard error and Tobit regression approaches produced smaller standard error estimates and subsequently significant results. Hence, we would like to further examine whether partially addressing the floor effects (i.e., the robust standard error approach) would be sufficient enough to obtain unbiased parameter estimates and standard errors, or only fully addressing the floor effects (i.e., Tobit regression) would result in unbiased estimates and standard errors.

THE SIMULATION STUDY

To further examine the impact of floor effects in multilevel data analysis, a Monte Carlo simulation study was conducted. Using the Monte Carlo routine in M*plus* version 7.31 (Muthén and Muthén, 1998-2015), data with floor effects were generated. Next, the simulated data were analyzed using the three comparative methods: (a) the maximum likelihood (ML) based multilevel regression model without addressing the floor effects, (b) the robust standard error approach (i.e., the ML based multilevel regression model with robust standard error estimator) only partially addressing the floor effects, and (c) the multilevel Tobit model which fully addressing the floor effects by defining the outcome variable as a left-censored variable.

Data Generation

Data were simulated based on a basic two-level random intercept model which was a commonly used multilevel structural equation model and could be fitted with the M*plus* Type=Twolevel routine. Floor effects in the outcome variable were considered. The population model for data generation was as follows. The fixed effects parameter vector (γ_{00}, γ_{10}) represented the grand mean and slope. ϕ represented the between-level variance and σ_i^2 was the within-level residual variance.

$$\text{Level 1}: \mathrm{Y_{ij}} = \beta_{0j} + \beta_{1j}\mathrm{X1_{ij}} + \mathrm{e_{ij}} \tag{15}$$

$$\text{Level 2}: \beta_{0j} = \gamma_{00} + \gamma_{01}\mathrm{X2_j} + \mathrm{U_{0j}} \tag{16}$$

$$\beta_{1j} = \gamma_{10} \tag{17}$$

$$\mathrm{U_{0j}} \sim N(0, \phi) \tag{18}$$

$$\mathrm{e_{ij}} \sim N(0, \sigma_i^2). \tag{19}$$

Population parameters used to generate the data are as follows: the variances of X1 and X2 were both 1. The means of X1 and X2 were set to be zero. The within-level residual variance ϕ and the between-level residual variance σ_i^2 were set to equal 1 and 0.5, respectively. The between-level mean of Y was set to be 1. The parameter vector (γ_{10}, γ_{01}) was set to be (0.75, 0.50). The outcome Y_{ij} was simulated with different proportions of floor data, which was detailed in a later section. Sample size was 1,000. Five hundred replications were generated for each simulation condition.

Regarding the different proportions of floor data in the outcome Y, six conditions (i.e., 0, 5, 10, 15, 20, and 25%) were considered, with 0% representing no floor effects and 25% representing the most floor effects. In the study by Wang et al. (2008), the authors used different ceiling thresholds to manipulate different ceiling proportion conditions in studying ceiling effects. In this study, we adopted their approach and varied the left-censoring points to create different proportions of floor data. The floor proportions and floor thresholds are presented in **Table 4**. The proportion of floor data increased as floor thresholds increased. **Figure 2** displays the corresponding distributions of the six simulated data sets with different proportions of floor data in the outcome variable. In the 0% floor data condition, the data was shown to be normally distributed. The 0% proportion condition served as the baseline condition. When the proportion of floor data increased (e.g., 5–25%), scores increasingly stacked on the lower end and the data distribution further shifted to the left (or more right skewed).

TABLE 4 | Floor proportions with different floor thresholds.

Proportions of floor data (%)	Floor thresholds	Mean (*SD*)	Score range
0	No floor	0.86 (1.60)	(−4.62, 5.60)
5	−1.50	1.04 (1.47)	(−1.50, 5.61)
10	−0.95	1.08 (1.41)	(−0.95, 5.61)
15	−0.60	1.12 (1.35)	(−0.60, 5.61)
20	−0.25	1.18 (1.28)	(−0.25, 5.61)
25	0.03	1.25 (1.21)	(0.03, 5.61)

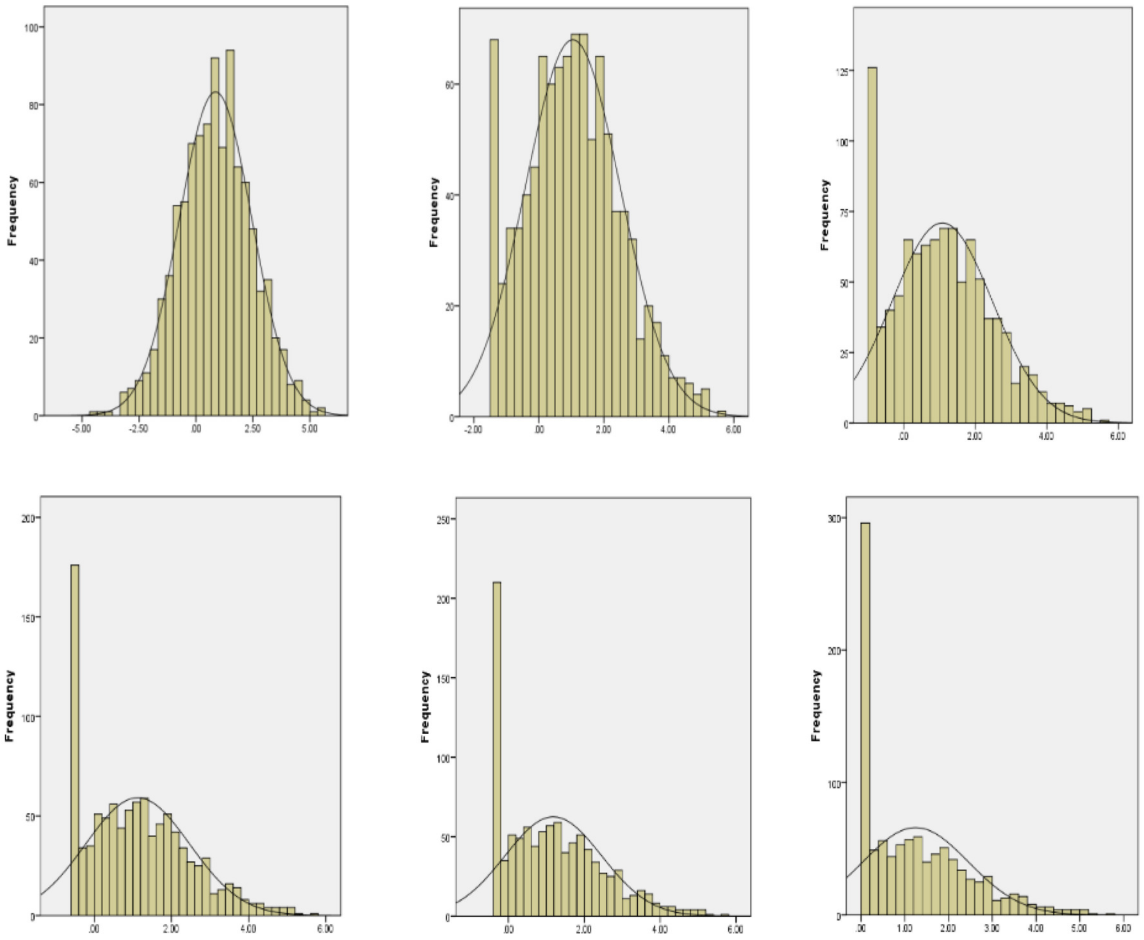


FIGURE 2 | Distribution of six simulated data sets showing different proportions of floor data (from 0 to 25% floor data).

Data Analysis

The simulated data were analyzed using three comparative methods as described previously. While the regular multilevel regression ignore the floor effects, robust standard error approach and multilevel Tobit regression model focuses on the floor effects, with the former one partially addresses the floor effects and the later one fully addresses the floor effects.

Simulation Results

The results are summarized across all 3,000 replications with respect to different methods in dealing with the increasing proportions of floor data in the outcome variables. Results of the relative bias in parameter estimates and standard errors for the three methods (i.e., regular multilevel regression, robust standard error approach, and multilevel Tobit regression model) are presented in **Tables 5**, **6**, respectively. Relative bias in parameter estimates was given as $(\hat{\theta} - \theta)/\theta$ where $\hat{\theta}$ represented the average estimate and θ was the corresponding population value. Similarly, relative bias in standard error estimates was given as $(\hat{\sigma} - \sigma)/\sigma$ where $\hat{\sigma}$ represented the average standard error estimate and σ was the corresponding population value. The differences in coverage values, statistical powers, type I error rates, and model fit statistics (i.e., CFI, RMSEA, and SRMR) were negligible across the three methods. Next, the relative biased in parameter estimates and the corresponding standard error estimates were discussed for the three different methods, namely, (1) the ML-based regular multilevel analysis without addressing the floor effects, (2) the robust standard error approach with partially addressing the floor effects (as correction for the non-normality in the floor data), and (3) the multilevel Tobit regression approach with fully addressing the floor effects in data.

Table 5 presents the relative bias in parameter estimates comparing the three methods. The regular multilevel analysis without addressing floor effects led to the underestimation in parameter estimates and the underestimation became substantial as the proportion of floor data increased. The robust standard error approach which only partially addressing the floor data yielded similar results as the regular multilevel analysis. The reason was that the robust standard error approach only corrected for standard error estimates rather than parameter estimates when the normality assumption was violated. As

TABLE 5 | Relative bias in parameter estimates comparing three comparative methods.

Proportions of floor data (%)	Parameter estimates					
	γ_{10}			γ_{01}		
	Multilevel model floors ignored	Robust standard error	Multilevel model floors considered	Multilevel model floors ignored	Robust standard error	Multilevel model floors considered
0	−0.00	−0.00	−0.00	−0.01	−0.01	−0.01
5	−0.05	−0.05	0.00	−0.05	−0.05	0.00
10	−0.10	−0.10	0.00	−0.10	−0.10	0.00
15	−0.15	−0.15	0.00	−0.14	−0.14	0.00
20	−0.20	−0.20	0.00	−0.20	−0.20	0.00
25	−0.26	−0.26	0.00	−0.26	−0.26	0.00

shown in **Table 5**, only the multilevel Tobit regression approach yielded the unbiased parameter estimates (i.e., $(\hat{\theta} - \theta)/\theta = 0$). This approach fully addressed the floor effects by treating the outcome variable as a left-censored variable. In this case, parameter estimates recovered well in the multilevel Tobit regression approach regardless the proportion of the floor data.

Table 6 summarizes the relative bias in standard errors using the three different methods. There was a clear pattern showing a systematic underestimation of standard errors when floor effects were ignored in regular multilevel analysis. The biases in the robust standard error approach were either similar or smaller than the ones in the regular multilevel analysis approach, and the standard errors were persistently underestimated. Furthermore, as the proportion of floor data increased, the biases tended to be larger. As shown in **Table 6**, among the three methods, the multilevel Tobit regression approach yielded the smallest bias. Given that most values were around 0.01 and the pattern was stable regardless the proportions of floor data, the degree of underestimation in the standard error estimates for the multilevel Tobit regression approach was negligible. This simulation demonstrated the importance of fully addressing the floor effects in multilevel data and the advantage of using the multilevel Tobit regression over the other methods when analyzing potential floor effects in the data.

DISCUSSION

This study highlighted the impact of floor effects when the PPVT-4 and the EVT-2 used with a culturally and linguistically diverse population of preschoolers by examining the impact of floor effects in data analysis, including in estimating parameters, the corresponding standard errors, and the tests of significance. Influences of floor effects in multilevel data analysis were investigated through an empirical example and a Monte Carlo simulation study.

Our findings suggest that some caution is warranted when interpreting findings from both PPVT-4 or the EVT-2, especially when these two tests are used with culturally, ethnically and linguistically and ethnical diverse groups. Given the standardized PPVT-4 and EVT-2 are both normed based on the predominantly White, middle-class, English-speaking American samples (Qi et al., 2003), ethnically, culturally and linguistically diverse groups may perform inferiorly relative to the normed sample especially among non-English or non-standard English speaking children. Samples in which 15% or more score at or near to the lowest level in the instruments measurement range may indicate the potential existence of the floor effects which may impact analyses when using traditional analytic methods (Lim et al., 2015). Given the culturally, linguistically and ethnically diverse populations have been shown to perform poorly on the PPVT-4 and the EVT-2 (e.g., Champion et al., 2003; Gonzalez et al., 2015), researchers should attend to with the potential problem of floor effects when using these measures. With increasing populations of language-minority populations (e.g., ELLs, DLLs), considering floor effects in measures warrants close attention.

In this study, it was demonstrated that when analyzing data shown to have floor effects, analytical methods insufficiently addressing the floor effects can lead to misleading results and interpretations First, when investigating the shared-reading intervention effects with a sample of Mexican-American preschoolers enrolled in the randomized clinical study using the PPVT-4 and EVT-2, failing to consider the impact of floor effects led to non-significant effects and quite possibly, misleadingly underestimated the impact of the intervention. Outcomes in this study supported previous findings suggesting that (Hessling et al., 2004), floor effects may undermine the true effects of an intervention, especially among linguistically and culturally diverse populations.

Furthermore, results from the simulation study showed that ignoring floor effects resulted in substantial bias in both parameter estimates and standard errors estimates, and erroneous significance tests. These findings are important and support previous research. McBee (2010) stated conventional statistical methods (e.g., ANOVA, linear regression) produced biased estimates when floor effects were present. Wang et al. (2008) also pointed out the consequence of biased parameter estimates due to the ceiling effects or floor effects. In our study, the two insufficient approaches, namely, the regular multilevel regression ignoring floor effects and the robust standard error approach which only partially addressing floor effects, produced the same parameter estimates. However, the robust standard

TABLE 6 | Relative bias in standard error estimates comparing three comparative methods.

Proportions of floor data (%)	Standard error estimates					
	$SE\gamma_{10}$			$SE\gamma_{01}$		
	Multilevel model floors ignored	Robust standard error	Multilevel model floors considered	Multilevel model floors ignored	Robust standard error	Multilevel model floors considered
0	0.00	−0.01	−0.01	−0.03	−0.04	−0.04
5	−0.07	−0.09	−0.03	−0.06	−0.09	−0.03
10	−0.14	−0.14	−0.05	−0.10	−0.44	0.00
15	−0.16	−0.13	−0.03	−0.16	−0.18	0.01
20	−0.21	−0.15	−0.02	−0.21	−0.23	−0.01
25	−0.27	−0.17	−0.02	−0.27	−0.27	−0.01

error approach produced less but still biased standard error estimates due to the "robust" correction. Multilevel Tobit regression was the only method that recovered all the parameter and standard error estimates very well. The multilevel Tobit regression model treated the outcome variables with floor effects as left-censored variables. In other words, scores on the very low end that could not be accurately measured due to the restricted range of the standardized assessments were treated as being left-censored. The Monte Carlo study showed that the multilevel Tobit regression effectively handled floor data. For example, even as low as only 5% of floor data could lead to biased results if floor effects were not adequately and fully addressed. Parameter estimates and standard error estimates were underestimated. The magnitude of the bias became larger as the proportion of floor data increased. Taken together, researchers should consider using the multilevel Tobit regression model to analyze the data with potential floor effects.

Finally, in order to examine floor data, graphs (e.g., histograms) can be easily and effectively used to illustrate whether a substantial proportion of scores stack at the lower end of the distribution. If there is a large percentage of very low scores in their data, researchers should consider the presence of floor effects. Again, as demonstrated in both empirical example and simulation studies, it is important to fully address the floor effects with adequate method, the Tobit regression given that insufficiently addressing the floor effects can result in biased parameter estimates and standard errors, which in turn, can lead to incorrect statistical inferences.

In summary, researchers need to be aware and cautious of the potential for floor effects when analyzing data from ethnically, culturally and linguistically diverse children accessed by the PPVT-4 and the EVT-2. A potential indicator for floor data is the disproportional representation of scores at the lower end of the distribution of the measured scores. Ignoring floor effects can lead to biased parameter estimates and standard errors, and quite possibly serious misleading inferences. It is thereby important for applied researchers who use standardized vocabulary tests with diverse populations to examine their data for floor effects and consider alternatives to the traditional data analysis methods which without fully addressing the floor effects. For modeling outcome variables with floor data, multilevel Tobit regression model is the recommended method for analyzing this type of data.

AUTHOR CONTRIBUTIONS

LZ initiated the design of the study and presented the work in the 2016 American Psychological Association Annual Convention. LZ and JG wrote the paper. Research data are taken from JG's Project Words of Oral Reading and Language Development (WORLD) efficacy studies (R350A110638: 2011-2014). LZ performed the Monte Carlo modeling.

REFERENCES

Baraldi, A. N., and Enders, C. K. (2010). "Missing data methods," in *The Oxford Handbook of Quantitative Methods*, ed T. Little (New York, NY: Oxford University Press), 635–664.

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *Brit. J. Math. Stat. Psychol.* 37, 62–83. doi: 10.1111/j.2044-8317.1984.tb00789.x

Brown, T. A. (2006). *Confirmatory Factor Analysis for Applied Research*. New York, NY: Guilford.

Bucik, N., and Bucik, V. (2003). Peabody slikovni besedni test (Peabody Picture Vocabulary Test, PPVT-III): Njegove merske lastnosti in uporabna vrednost. = Peabody Picture Vocabulary Test (PPVT-III): psychometric properties and significance for application. *Horiz. Psychol.* 12, 91–108. Retrieved from: http://www.dlib.si/details/URN:NBN:SI:DOC-QM60VCCW

Buly, M. R. (2005). Leaving no American Indian/Alaska Native behind: identifying reading strengths and needs. *J. Am. Ind. Educ.* 44, 28–52.

Capovilla, F. C., and Capovilla, A. G. S. (1997). Desenvolvimento lingüístico na criança dos dois aos seis anos: tradução e estandardização do Peabody Picture Vocabulary Test de Dunn & Dunn. E Da Language Development Survey de Rescorla. *Ciência Cognitiva Teoria Pesquisa e Aplicação* 1, 353–380.

Catts, H. W., Petscher, Y., Schatschneider, C., Bridges, M. S., and Mendoza, K. (2009). Floor effects associated with universal screening and their impact on the early identification of reading disabilities. *J. Learn. Disabil.* 42, 163–176. doi: 10.1177/0022219408326219

Champion, T. B., Hyter, Y. D., McCabe, A., and Bland-Stewart, L. M. (2003). "A matter of vocabulary" performances of low-income African American Head Start children on the Peabody Picture Vocabulary Test—III. *Commun. Disord. Q.* 24, 121–127. doi: 10.1177/15257401030240030301

Cox, D. R., and Oakes, D. (1984). *Analysis of Survival Data*. London: Chapman and Hall.

DeAvila, E. A., and Duncan, S. E. (2000). *PreLAS2000: English and Spanish Technical Notes.* Monterey, CA: CTB/McGraw-Hill.

Dunn, L. M., and Dunn, D. M. (2007). *Peabody Picture Vocabulary Test, 4th Edn.* Minneapolis, MN: NCS Pearson.

Finney, S. J., and DiStefano, C. (2006). "Nonnormal and categorical data in structural equation models," in *Structural Equation Modeling: A Second Course*, eds G. R. Hancock and R. O. Mueller (Greenwich, CT: Information Age), 269–314.

Gonzalez, J., Pollard-Durodola, S., Saenz, L., Soares, D., Davis, H., Resendez, N., et al. (2015). Spanish and English early literacy profiles of preschool Latino English language learner children. *Early Educ. Dev.* 27, 513–531. doi: 10.1080/10409289.2015.1077038

Haitana, T., Pitama, S., and Rucklidge, J. J. (2010). Cultural biases in the peabody picture vocabulary test-III: testing tamariki in a New Zealand sample. *J. Psychol.* 39, 24–34.

Hessling, R. M., Schmidt, T. J., and Traxel, N. M. (2004). "Floor effect," in *Encyclopedia of Social Science Research Methods*, eds M. S. Lewis-Beck, A. Bryman, and T. F. T. Liao (Thousand Oaks, CA: SAGE Publications, Inc.), 390–391.

Horton-Ikard, R., and Ellis Weismer, S. (2007). A preliminary examination of vocabulary and word learning in African American toddlers from middle and low socioeconomic status homes. *Am. J. Speech Lang. Pathol.* 16, 381–392. doi: 10.1044/1058-0360(2007/041)

Hox, J. J., Maas, C. J., and Brinkhuis, M. J. (2010). The effect of estimation method and sample size in multilevel structural equation modeling. *Stat. Neerl.* 64, 157–170. doi: 10.1111/j.1467-9574.2009.00445.x

Iachina, M., and Iachina, N. (2012). Measuring reliable change of emotional and behavioural problems in children. *Psychiatry Res.* 200, 867–871. doi: 10.1016/j.psychres.2012.06.023

Ji, C., Yao, D., Chen, W., Li, M., and Zhao, M. (2014). Adaptive behavior in Chinese children with Williams syndrome. *BMC Pediatr.* 14:90. doi: 10.1186/1471-2431-14-90

Keeley, J., Keeley, T., English, J., and Irons, A. M. (2013). Investigating halo and ceiling effects in student evaluations of instruction. *Educ. Psychol. Meas.* 73, 440–457. doi: 10.1177/0013164412475300

Kim, Y., Chang, H., Lim, S., and Bak, H. (1995). *Geurim Eohyuryeok Geomsa [Picture Vocabulary Test]*. Seoul: Seoul Community Rehabilitation Center.

King, G., and Roberts, M. E. (2015). How robust standard errors expose methodological problems they do not fix, and what to do about it. *Polit. Anal.* 23, 159–179. doi: 10.1093/pan/mpu015

Klein, J. P., and Moeschberger, M. L. (1997). *Survival Analysis: Techniques for Censored and Truncated Data.* New York, NY: Springer. doi: 10.1007/978-1-4757-2728-9

Kleinbaum, D. G., and Klein, M. (2005). *Survival Analysis: A Self-Learning Text.* New York, NY: Springer-Verlag.

Laing, S. P., and Kamhi, A. (2003). Alternative assessment of language and literacy in culturally and linguistically diverse populations. *Lang. Speech Hear. Serv. Sch.* 34, 44–55. doi: 10.1044/0161-1461(2003/005)

Lim, C. R., Harris, K., Dawson, J., Beard, D. J., Fitzpatrick, R., and Price, A. J. (2015). Floor and ceiling effects in the OHS: an analysis of the NHS PROMs data set. *BMJ Open* 5:e007765. doi: 10.1136/bmjopen-2015-007765

Long, J. S. (1997). *Regression Models for Categorical and Limited Dependent Variables: Advanced Quantitative Techniques in the Social Sciences*. Thousand Oaks, CA: Sage Publications.

Lonigan, C. J., Farver, J. M., Nakamoto, J., and Eppe, S. (2013). Developmental trajectories of preschool early literacy skills: a comparison of language-minority and monolingual-English children. *Dev. Psychol.* 13, 1–15. doi: 10.1037/a0031408

Maas, C. J., and Hox, J. J. (2004). Robustness issues in multilevel regression analysis. *Stat. Neerl.* 58, 127–137.

McBee, M. (2010). Modeling outcomes with floor or ceiling effects: an introduction to the tobit model. *Gifted Child Q.* 54, 314–320. doi: 10.1177/0016986210379095

McCabe, A., and Champion, T. B. (2010). A matter of vocabulary II: low-income African American children's performance on the expressive vocabulary test. *Commun. Disord. Q.* 31, 162–169. doi: 10.1177/1525740109344218

Muthén, B. (1989). Tobit factor analysis. *Brit. J. Math. Stat. Psychol.* 42, 241–250.

Muthén, B. (1990). *Means and Covariance Structure Analysis of Hierarchical Data.* UCLA Statistics Series. Los Angeles, CA: Department of Statistics Ucla.

Muthén, B., and Asparouhov, T. (2010). "Beyond multilevel regression modeling: multilevel analysis in a general latent variable framework," in *Handbook of Advanced Multilevel Analysis*, eds J. Hox and J. K. Roberts (New York, NY: Taylor and Francis), 15–40.

Muthén, L. K., and Muthén, B. O. (1998-2015). *Mplus User's Guide 7th Edn*. Los Angeles, CA: Muthén & Muthén.

Nevitt, J., and Hancock, G. R. (2001). Performance of bootstrapping approaches to model test statistics and parameter standard error estimation in structural equation modeling. *Struct. Equat. Model.* 8, 353–377. doi: 10.1207/S15328007SEM0803_2

Pae, H. K., Greenberg, D., and Morris, R. D. (2012). Construct validity and measurement invariance of the Peabody Picture Vocabulary Test – III Form A. *Lang. Assess. Q.* 9, 152–171. doi: 10.1080/15434303.2011.613504

Pakendorf, C., and Alant, E. (1997). Culturally valid assessment tools: Northern Sotho translation of the Peabody Picture Vocabulary Test – Revised. *South Afr. J. Commun. Disord.* 44, 3–12.

Pollard-Durodola, S. D., Gonzalez, J. E., Saenz, L., Soares, D., Resendez, N., Kwok, O., et al. (2016). The effects of content-related shared book reading on the language development of preschool dual language learners. *Early Child. Res. Q.* 36, 106–121. doi: 10.1016/j.ecresq.2015.12.004

Proust-Lima, C., Dartigues, H., and Dartigues, H. (2011). Misuse of the linear mixed model when evaluating risk factors of cognitive decline. *Am. J. Epidemiol.* 174, 1077–1088. doi: 10.1093/aje/kwr243

Qi, C. H., Kaiser, A. P., Milan, S. E., Yzquierdo, Z., and Hancock, T. B. (2003). The performance of low-income, African American children on the Preschool Language Scale−3. *J. Speech Lang. Hear. Res.* 46, 576–590. doi: 10.1044/1092-4388(2003/046)

Raudenbush, S. W., and Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods, 2nd Edition*. Newbury Park, CA: Sage.

Restrepo, M. A., Schwanenflugel, P. J., Blake, J., Neuharth-Pritchett, S., Cramer, S. E., and Ruston, H. P. (2006). Performance on the PPVT–III and the EVT: applicability of the measures with African American and European American preschool children. *Lang. Speech Hear. Serv. Sch.* 37, 17–27. doi: 10.1044/0161-1461(2006/003)

Rock, D. A., and Stenner, A. J. (2005). Assessment issues in the testing of children at school entry. *Fut. Child.* 15, 15–34. doi: 10.1353/foc.2005.0009

Stockman, I. J. (2000). The new Peabody Picture Vocabulary Test-III: an illusion of unbiased assessment? *Lang. Speech Hear. Serv. Schools* 31, 340–353. doi: 10.1044/0161-1461.3104.340

Terry, N. P., Mills, M. T., Bingham, G. E., Mansour, S., and Marencin, N. (2013). Oral narrative performance of African American prekindergartners who speak nonmainstream American English. *Lang. Speech Hear. Serv. Sch.* 44, 291–305. doi: 10.1044/0161-1461(2013/12-0037)

Theriault-Whalen, C. M., and Dunn, L. M. (1993). *Echelle de Vocabulaire en Images.* Richmond Hill, ON: Peabody, Psyscan Corporation.

Thernstrom, A. (2002). "The racial gap in academic achievement," in *Beyond the Color Line: New Perspectives on Race and Ethnicity in America, 1st Edn.*, eds A. Thernstrom and S. Thernstrom (Stanford CA: Hoover Institution Press), 259–276.

Thomas-Tate, S., Washington, J., Craig, H., and Packard, M. (2006). Performance of African American preschool and kindergarten students on the expressive vocabulary test. *Lang. Speech Hear. Serv. Schools* 37, 143–149. doi: 10.1044/0161-1461(2006/016)

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica* 26, 24–36. doi: 10.2307/1907382

Twisk, J., and Rijmen, F. (2009). Longitudinal tobit regression: a new approach to analyze outcome variables with floor or ceiling effects. *J. Clin. Epidemiol.* 62, 953–958. doi: 10.1016/j.jclinepi.2008.10.003

Ueno, K., Utsuo, T., and Iinaga, K. (1991). *PVT Kaiga Goi Hattatsu Kensa [PVT Picture Vocabulary Development Test]*. Tokyo: Chiba Test Center.

Wang, L. J., Zhang, Z. Y., McArdle, J. J., and Salthouse, T. A. (2008). Investigating ceiling effects in longitudinal data analysis. *Multivar. Behav. Res.* 43, 476–496. doi: 10.1080/00273170802285941

Washington, J., and Craig, H. K. (1992). Performances of low-income African American preschool and kindergarten children on the Peabody Picture Vocabulary Test–revised. *Lang. Speech Hear. Serv. Sch.* 23, 329–333. doi: 10.1044/0161-1461.2304.329

Washington, J., and Craig, H. K. (1999). Performances of at-risk, African American preschoolers on the Peabody Picture Vocabulary Test–III. *Lang. Speech Hear. Serv. Sch.* 30, 75–82. doi: 10.1044/0161-1461.3001.75

Whitaker, S. (2005). The use of the WISC-III and the WAIS-III with people with a learning disability: three concerns. *Clin. Psychol. Forum* 50, 37–40.

Whitaker, S. (2008). Intellectual disability: a concept in need of revision. *Br. J. Dev. Disabil.* 54, 3–9. doi: 10.1179/096979508799103350

Whitaker, S. (2010). Error in the estimation of intellectual ability in the low range using the WISC-IV and WAISIII. *Pers. Individ. Dif.* 48, 517–521. doi: 10.1016/j.paid.2009.11.017

Whitaker, S. (2012). Review of the WAIS-IV - the measurement of low IQ with the WAIS-IV: a critical review. *Clin. Psychol. Forum* 45–48.

Whitaker, S., and Gordon, S. (2012). Floor effects on the WISC-IV. *Int. J. Dev. Disabil.* 58, 111–119. doi: 10.1179/2047387711Y.0000000012

Williams, K. T. (2007). *Expressive Vocabulary Test, 2nd Edn*. Circle Pines, MN: AGS Publishing.

Winship, C., and Mare, R. D. (1984). Regression models with ordinal variables. *Am. Sociol. Rev.* 49, 512–525. doi: 10.2307/2095465

Yuan, K. H., and Bentler, P. M. (2000). "Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data," in *Sociological Methodology*, eds M. E. Sobel and M. P. Becker (Washington, DC: ASA), 165–200.

Yuan, K.-H., Bentler, P. M., and Zhang, W. (2005). The effect of skewness and kurtosis on mean and covariance structure analysis: the univariate case and its multivariate implication. *Sociol. Methods Res.* 34, 249–258. doi: 10.1177/0049124105280200

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

APPENDIX A

Input Specifications in Mplus for the Multilevel Tobit Regression Model

```
TITLE: Syntax for a multilevel tobit model
DATA: FILE IS Floor.dat;
VARIABLE: NAMES are sid tid x1-x7 evt1 evt2 evts1 evts2 x8-x14 ppvt1 ppvt2 ppvts1 ppvts2;
USEVARIABLES are x1-x7 evts1 evts2 x8-x14;
!outcome variable (ects2) with floor effects
CENSORED ARE evts2 (b);
CLUSTER = tid;
WITHIN = x1-x7 evts1;
BETWEEN = x8-x14;
MISSING is all (9999);
ANALYSIS:
TYPE = twolevel;
!robust standard error
ESTIMATOR = MLR;
MODEL:
%WITHIN%
evts2 ON x1-x7 evts1;
%BETWEEN%
evts2 ON x8-x14;
OUTPUT: TECH1 standardized
```

frontiers in Psychology

ORIGINAL RESEARCH
published: 22 September 2017
doi: 10.3389/fpsyg.2017.01464

A Solution to Modeling Multilevel Confirmatory Factor Analysis with Data Obtained from Complex Survey Sampling to Avoid Conflated Parameter Estimates

Jiun-Yu Wu[1]*, *John J. H. Lin*[2], *Mei-Wen Nian*[1] and *Yi-Cheng Hsiao*[1]

[1] *Institute of Education, National Chiao Tung University, Hsinchu, Taiwan,* [2] *Office of Institutional Research, National Central University, Taoyuan, Taiwan*

OPEN ACCESS

Edited by:
Holmes Finch,
Ball State University, United States

Reviewed by:
Hsien-Yuan Hsu,
University of Texas Health Science Center at Houston, United States
Jung Yeon Park,
Columbia University, United States
Ke-Hai Yuan,
University of Notre Dame, United States

***Correspondence:**
Jiun-Yu Wu
jiunyu.rms@gmail.com;
jiunyuwu@g2.nctu.edu.tw

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *14 March 2017*
Accepted: *15 August 2017*
Published: *22 September 2017*

Citation:
Wu J-Y, Lin JJH, Nian M-W and Hsiao Y-C (2017) A Solution to Modeling Multilevel Confirmatory Factor Analysis with Data Obtained from Complex Survey Sampling to Avoid Conflated Parameter Estimates. Front. Psychol. 8:1464. doi: 10.3389/fpsyg.2017.01464

The issue of equality in the between-and within-level structures in Multilevel Confirmatory Factor Analysis (MCFA) models has been influential for obtaining unbiased parameter estimates and statistical inferences. A commonly seen condition is the inequality of factor loadings under equal level-varying structures. With mathematical investigation and Monte Carlo simulation, this study compared the robustness of five statistical models including two model-based (a true and a mis-specified models), one design-based, and two maximum models (two models where the full rank of variance-covariance matrix is estimated in between level and within level, respectively) in analyzing complex survey measurement data with level-varying factor loadings. The empirical data of 120 3rd graders' (from 40 classrooms) perceived Harter competence scale were modeled using MCFA and the parameter estimates were used as true parameters to perform the Monte Carlo simulation study. Results showed maximum models was robust to unequal factor loadings while the design-based and the miss-specified model-based approaches produced conflated results and spurious statistical inferences. We recommend the use of maximum models if researchers have limited information about the pattern of factor loadings and measurement structures. Measurement models are key components of Structural Equation Modeling (SEM); therefore, the findings can be generalized to multilevel SEM and CFA models. Mplus codes are provided for maximum models and other analytical models.

Keywords: multilevel confirmatory factor analysis, design-based approach, model-based approach, maximum model, level-varying factor loadings, complex survey sampling, measurement

INTRODUCTION

Multilevel Confirmatory Factor Analysis (MCFA) extends the power of Confirmatory Factor Analysis (CFA) to accommodate the complex survey data with the estimation of the level-specific variance components and the respective measurement models. Complex survey data are obtained through cluster sampling or multistage sampling, where a few individuals within a class/household or the entire class/family are selected. This type of sampling scheme is likely to result in non-independent observations with within-cluster dependency (Skrondal and Rabe-Hesketh, 2007). If the dependent data are analyzed through the traditional approaches which assume independent observations, "incorrect parameter estimates, standard errors, and inappropriate fit statistics may be obtained" (du Toit and du Toit, 2008, p. 456).

Researchers has devoted their attention in discussing the influences of applying different multilevel modeling constructions on complex survey data (e.g., model-fit indices: Hsu et al., 2015; reliability measures: Geldhof et al., 2013; parameter estimates and statistical inferences: Wu and Kwok, 2012; longitudinal design: Wu et al., 2014). Among the research designs in these studies, the issue of inequality in the between- (i.e., the higher level or cluster level) and within-level (i.e., the lower level or individual level) structure in complex survey data has been proven to be influential for obtaining unbiased parameter estimates along with their consistent statistical inferences. Compared to inequality of level structures in multilevel models, a less addressed condition is that the true model did have the same factor structure at both levels while the magnitudes and statistical significance of the factor loadings varied across levels and varied within the levels, which occurred frequently in empirical research (e.g., Dyer et al., 2005; Klangphahol et al., 2010).

For example, Dyer et al. (2005) applied MCFA to study organizational leadership at the individual and societal level and obtained a common factor consisting of five items of being "formal," "habitual," "cautious," "procedural," and "ritualistic." The five items loaded much stronger onto the single factor at the between level (i.e., societal level) than at the within level (individual level), which supported the belief that this leadership scale operates mainly at the societal level. Based on this finding, Dyer et al. (2005) suggested that a three-item factor (discarding two trivial items with small factor loadings) instead of a five-item factor should be used if the interest of leadership study is at the individual level. Dyer et al.'s suggestion capitalized on the importance of specifying an optimal measurement model with complex survey data in terms of both model structure and sizes of factor loadings to obtain correct statistical and practical interpretations in scale development.

From the factor analysis point of view, items with variance explained smaller than 20% or standardized factor loadings less than 0.45 would be considered as low communality (EFA: MacCallum et al., 1999; CFA: Meade and Bauer, 2007). From a measurement point of view, items with standardized factor loadings larger than 0.6 would exhibit better psychometric properties (Bagozzi and Yi, 1988; Kline, 2010). Failing to detect items with small factor loadings may lead to a misunderstanding that all items are equally important, causing researchers to investigate problems that are of little importance or little relevance to the intended measure.

Therefore, in this study, we performed a substantive-methodological synergy (Marsh and Hau, 2007) by applying different modeling strategies on simulated synthetic datasets with population parameters specified based on an empirical dataset to examine the robustness of model-based, design-based, and maximum models regarding their effectiveness and efficiency in producing unbiased parameter estimates and statistical inference for the measurement data obtained from complex survey sampling. Below we elaborated on the issues with modeling strategies and unequal factor loadings, followed by introduction to three modeling strategies on complex survey data.

Issues with Modeling Strategies and Unequal Factor Loadings

Traditionally, several multilevel modeling strategies can be applied to address data dependency in complex survey data (Heck and Thomas, 2008; Rabe-Hesketh and Skrondal, 2008; Hox, 2010; Snijders and Bosker, 2011). Specifying different structures for separate levels, namely a model-based approach, on complex survey data allows free estimation of level-specific parameters and enables the detection of possible inequality in parameter estimates. However, in reality, information or truth about the higher-level structure is rarely known without the support of theoretical evidence. If researchers jump into multilevel analysis without theoretical or empirical evidence, the correctness of the multilevel structure is at risk.

Alternatively, researchers can apply the design-based approach by specifying only an overall model for the complex survey data to infer their findings to the lower level sampling units, and using the robust standard error estimator (Huber, 1967; White, 1980) to correct for the bias in standard error of the fixed effects (Muthén and Satorra, 1995). The design-based approach has been proved to yield satisfying analytic results only when the complex survey data meet the assumption of equal structures in both between- and within-levels (Wu and Kwok, 2012). In addition to design-based and model-based approaches, a possible alternative for analyzing multilevel data is through the use of maximum models (Hox, 2002, 2010; Wu and Kwok, 2012), where a saturated between-level model is estimated and can be used to focus on a specific level of analysis.

To examine the robustness of reliability measures on complex survey data, Geldhof et al. (2013) used MCFA and single-level CFA (i.e., without taking data dependency into consideration) on the simulated multilevel datasets, where the between and within levels had exactly the same factor loadings but with different high and low reliability across levels using the average item ICC as a dependency measure. Their study findings suggested that single-level CFAs cannot yield the actual scale reliability unless the true reliabilities are identical at each level. Moreover, in the simulation study, they postulated that the true MCFA model had the same factor loadings within and across levels, i.e., the between and within level model were identical in terms of magnitude of factor loadings and factor structures. Few studies have investigated the issue of inequality of factor loadings under equal factor structure within and across levels. Besides, systematic investigation on the performance of model-fit statistics, indices and information criteria, and the resulted parameter estimates with statistical inferences were not discussed in Geldhof et al. (2013).

Extending the simulation settings of Geldhof et al. (2013), we examined performance of different model specifications regarding the issues of inequality of factor loadings and different factor structures within and across levels. Distinct Cluster Numbers (CN), Cluster Sizes (CS) and ICCs were used in true model for the simulation settings of this study. The criterion variables include overall exact model fit chi-square test and various model fit indices, both fixed-effect and random-effect parameter estimates, their 95% coverage rate and empirical power, as well as the variance explained measure (R^2) and scale reliability (ρ).

Specifically, this study aims to examine the robustness of the three modeling strategies using five analytic models (i.e., MCFA, miss-specified MCFA, one-level design-based CFA, Max CFA with saturated Between level, Max CFA with saturated Within level) in testing the multilevel measurement data with unequal magnitudes of factor loadings. Of the factor loadings, some may be trivial or of little relevance in a practical sense at the individual level under equal level structures. In the following section, we provide a review of three multilevel modeling strategies.

Three Modeling Strategies on Complex Survey Data

Model- and Design-Based Strategies

The rationale for using multilevel models in analyzing complex survey data is to reflect the natural multistage sampling scheme (Muthén, 1994; Heck and Thomas, 2008). Researchers can do so by constructing the analytic model either to simultaneously calculate the lower- and higher-level parameter estimates which may have different values at each level or to adjust the standard errors of fixed effects. The model-based approach (e.g., MCFA technique) conforms to the actual multi-stage sampling scheme by specifying a level-specific model for each level of the data. In other words, for a two-level clustered sampling data, it specifies a between-level model that conforms to the level 2 structure (i.e., higher level) and a within-level model that conforms to the level 1 structure (i.e., lower level). Instead of constructing separate level models for multilevel data, the design-based approach analyzes the data with only one overall model and considers the sampling scheme by adjusting for the standard errors of the parameter estimates based on the sampling design. The adjustment is implemented using the robust standard error estimator (Huber, 1967; White, 1980) or sandwich-type variance estimator, a general name for alternative variance estimators. The sandwich-type variance estimator functions as an overall adjustment of the deviated standard error of parameter estimates due to extra data dependency along with the original statistical approach. This kind of relative variance estimators has been proposed to address data non-independence (i.e., data heteroskedasticity) more directly in CFAs (Muthén and Satorra, 1995). The adjustment is a *post-hoc* process and is said to only affect the standard errors, not the parameter estimates (Hardin and Hilbe, 2007).

In a simulation study, Muthén and Satorra (1995) showed that under the same model structure for all data levels, these two approaches performed equally well for complex survey data. Compared to the model-based approach, the design-based approach is used more frequently by researchers in the applied areas (Rebollo et al., 2006; Róbert, 2010; Roberts et al., 2010; Rosenthal and Villegas, 2010; Wu et al., 2010; Brook et al., 2011; Martin et al., 2011; Wu, 2015, 2017) because it only requires a single model specification and often researchers were interested in examining the lower level (i.e., the within-level) model with the most sampling units.

Despite the simplicity of the model's specifications, the design-based approach for complex survey data is built upon the assumption of the same level-varying structures (Muthén and Satorra, 1995; Wu and Kwok, 2012). However, this assumption is often violated in empirical research when researchers examine the level-specific structures of their multilevel dataset (e.g., Wilhelm and Schoebi, 2007). Inequality in the between- and within-level structures leads to conflated estimations of the fixed and random effects if the design-based approach is used (Wu and Kwok, 2012). What's more, in the current study, we posit that if the same magnitude and significance of factor loadings do not hold at different levels under same level structures, inequality of the between- and within-level factor loadings may also cause potential problem with the design-based approach. In the case of Dyer et al. (2005), if the authors had used the design-based approach for their procedural leadership analysis, they would obtain the design-based estimates which would have been contaminated with information from both the between- and within-level models. Thus, they would have no idea of the larger factor loadings at the societal level and may not be able to detect the two trivial items at the individual level. From a practical perspective, researchers would falsely conclude the scale is a valid measurement for the research question related to the individual participant. In addition, the estimation of the overall model parameters and the scale reliability measures may be questionable to infer the individual-level characteristics. However, the issue of inequality of factor loadings between and within the levels has rarely been systematically examined in previous studies.

Maximum Model

Another feasible modeling strategy for complex survey data is called the "maximum model," (Hox, 2002, 2010; Wu and Kwok, 2012) where a saturated model in specific level (usually the higher-level) is built by estimating the full rank of between-level variance-covariance matrix with the consumption of all available degrees of freedom. This maximum model technique was firstly suggested by researchers (e.g., Hox, 2002; Stapleton, 2006; Yuan and Bentler, 2007) as the baseline model for constructing multilevel analysis with theoretical evidence. Ryu and West (2009), on the other hand, examined the performance of level-specific fit indices using maximum modeling technique. More recently, Wu and Kwok (2012) found that the maximum model and correctly specified model-based approaches performed equally well for analyzing complex survey data regardless of equality in level structures whereas the design-based approach only produced satisfying fixed-effect estimates and standard error under equal within-/between-level structure scenarios. Compared to inequality of level structures, what is more commonly found in empirical measurement research is unequal magnitudes of factor loadings in different levels with the same number of factors. However, no study to date has systematically examined the consequences of miss-specifying multilevel models for a two-level CFA measurement data regarding the violation of equality of factor loadings. This study will focus on inequality of factor loadings within and across levels of MCFA to explore potential analytical problems.

In the SEM framework, analysts commonly use differential chi-square tests to conduct model comparison analysis with numerous completing models. However, this kind of test is only good for comparisons between nested models. Besides, the chi-square test statistic is easily influenced by large sample sizes (Yuan et al., 2007; Kline, 2010). Alternatively, information criteria statistics can be used for model comparison between nested and non-nested models (Sclove, 1987). By taking the model uncertainty into consideration, the information criteria overcome the above-mentioned difficulties (Bollen et al., 2014). In this study, besides commonly-used model-fit test statistics and indices, we discussed the performance of Akaike Information Criterion (AIC, Akaike, 1974), Bayesian/Schwartz Information Criterion (BIC, Schwarz, 1978), and the sample-size adjusted BIC (adj. BIC, Sclove, 1987) in assessing the different model specifications. Models with smaller AIC, BIC, or adjusted BIC would be considered a better fit to the designated dataset. Detailed discussion among these information criteria under the SEM framework can be found in Nylund et al. (2007) for Latent Class Analysis and Growth Mixture Modeling, and Bollen et al. (2014) for single-level SEM modeling. This study would add to the literature regarding the guideline of interpreting information criteria to construct measurement models for complex survey data under the SEM framework.

METHODS

Mathematical Investigation of Three SEM Techniques on Complex Survey Measurement Data

We provided the model specifications of the model-based, design-based, and maximum modeling approaches (with both saturated between-level model, and saturated within-level structure model) and their mathematical derivations to investigate the robustness of these modeling approaches in dealing with the inequality of factor loadings at between- and within-level models under equal factorial structures.

Using multilevel data drawn from a two-level multistage sampling strategy as an example, let us suppose that the G groups are randomly drawn from the target population at the first stage of sampling and that n_g participants are sampled within each group g at the second stage. We have a total of $N = \sum_{g=1}^{G} n_g$ participants. For each participant, P item responses (y_{pig}, p = 1,2,…,P) are gathered. We now have random vector of response variables $\mathbf{y}_{ig} = [y_{1ig}, y_{2ig}, \ldots, y_{pig}]_{1\times \mathrm{P}}$ for participant i (lower-level unit, i = 1,2,…,n_g) within group g (higher-level unit, g = 1,2,…,G).

For the gth group, the random matrix of observations may be arranged as follows:

$$\mathbf{y}_g = \begin{bmatrix} \mathbf{y}_{1g} \\ \mathbf{y}_{2g} \\ \vdots \\ \mathbf{y}_{n_g g} \end{bmatrix} = \begin{bmatrix} [y_{11g}\ y_{21g}\ \cdots\ y_{P1g}] \\ [y_{12g}\ y_{22g}\ \cdots\ y_{P2g}] \\ \vdots\ \vdots\ \cdots\ \vdots \\ [y_{1n_g g}\ y_{2n_g g}\ \cdots\ y_{Pn_g g}] \end{bmatrix}_{n_g \times P} \tag{1}$$

Analogous to the variance decomposition used in ANOVA analysis, the observation $\mathbf{y}_{ig}$ can be decomposed into its between-group component and within-group component, that is,

$$\mathbf{y}_{ig} = \mathbf{y}_{B\ldots g} + \mathbf{y}_{W.ig}, \forall i = 1, 2, \ldots, n_g, g = 1, 2, \ldots, G \tag{2}$$

where $\mathbf{y}_{B\ldots g}$ is the between-group component with $MVN\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}_\mathbf{B}\right)$ (i.e., multivariate normal distribution with grand mean $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}_\mathbf{B}$) and $\mathbf{y}_{W.ig}$ is the within-group component with $MVN\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}_\mathbf{W}\right)$. Typically, $\boldsymbol{\mu}_g$ is set as 0. The between-group components in different groups is set to be uncorrelated; that is, $Cov\left(\mathbf{y}_{B\ldots g}, \mathbf{y}_{B\ldots g'}\right) = \mathbf{0}, \forall g \neq g'$. Similarly, the correlation between different participants in different groups is also set to be zero (i.e., $Cov\left(\mathbf{y}_{W.ig}, \mathbf{y}_{W.i'g'}\right) = \mathbf{0}, \forall i \neq i'$ & $\forall g \neq g'$). Furthermore, the cross-level correlation between $\mathbf{y}_{B\ldots g}$ and $\mathbf{y}_{W.ig}$ is defined as uncorrelated.

Hence, the variance-covariance matrix of $\mathbf{y}_{ig}$may be decomposed into the combination of between-group and within-group variations, $Cov\left(\mathbf{y}_{ig}\right) = \boldsymbol{\Sigma}_B + \boldsymbol{\Sigma}_W$. Going a step further, to consider the MCFA model (i.e., the model-based approach), Equation (2) may be written as

$$\begin{aligned} \mathbf{y}_{ig} &= \mathbf{y}_{B\ldots g} + \mathbf{y}_{W.ig} \\ &= \boldsymbol{\mu} + \boldsymbol{\Lambda}_B \boldsymbol{\eta}_{B\ldots g} + \boldsymbol{\varepsilon}_{B\ldots g} + \boldsymbol{\Lambda}_W \boldsymbol{\eta}_{W.ig} + \boldsymbol{\varepsilon}_{W.ig} \end{aligned} \tag{3}$$

The between-group component $\mathbf{y}_{B\ldots g}$ is the combination of a product of factor loading matrix $\boldsymbol{\Lambda}_B$ and latent factor $\boldsymbol{\eta}_{B\ldots g} \sim MVN\left(\mathbf{0}, \boldsymbol{\Psi}_\mathbf{B}\right)$, and the unique vector $\boldsymbol{\varepsilon}_{B\ldots g} \sim MVN\left(\mathbf{0}, \boldsymbol{\Theta}_\mathbf{B}\right)$. The within-group component $\mathbf{y}_{W.ig}$ is the combination of a product of factor loading matrix $\boldsymbol{\Lambda}_W$ and latent factor $\boldsymbol{\eta}_{W.ig} \sim MVN\left(\mathbf{0}, \boldsymbol{\Psi}_\mathbf{W}\right)$, and the unique vector $\boldsymbol{\varepsilon}_{W.ig} \sim MVN\left(\mathbf{0}, \boldsymbol{\Theta}_\mathbf{W}\right)$. Random components were set to be orthogonal (i.e.,$\boldsymbol{\eta}_{B\ldots g} \perp \boldsymbol{\varepsilon}_{B\ldots g} \perp \boldsymbol{\eta}_{W.ig} \perp \boldsymbol{\varepsilon}_{W.ig}$).

Equation (3) specifies two sources of random variation for the observed variables, within-group (i.e., within-level) variation and between-group (i.e., between-level) variation to the nature of complex survey data, rather than just one overall random source. As a result, the variance-covariance matrix of $\mathbf{y}_{ig}$ may be further rewritten as

$$\begin{aligned} Cov\left(\mathbf{y}_{ig}\right) &= Cov\left(\boldsymbol{\mu}_B + \boldsymbol{\Lambda}_B \boldsymbol{\eta}_{B\ldots g} + \boldsymbol{\varepsilon}_{B\ldots g} + \boldsymbol{\mu}_W + \boldsymbol{\Lambda}_W \boldsymbol{\eta}_{W.ig} \right. \\ &\qquad \left. + \boldsymbol{\varepsilon}_{W.ig}\right) \\ &= Cov\left(\boldsymbol{\Lambda}_B \boldsymbol{\eta}_{B\ldots g} + \boldsymbol{\varepsilon}_{B\ldots g}\right) + Cov\left(\boldsymbol{\Lambda}_W \boldsymbol{\eta}_{W.ig} + \boldsymbol{\varepsilon}_{W.ig}\right) \\ &= \boldsymbol{\Lambda}_B \boldsymbol{\Psi}_B \boldsymbol{\Lambda}_B{}' + \boldsymbol{\Theta}_B + \boldsymbol{\Lambda}_W \boldsymbol{\Psi}_W \boldsymbol{\Lambda}_W{}' + \boldsymbol{\Theta}_W \qquad \text{(MCFA)} \end{aligned} \tag{4}$$

The variance covariance matrix of indicators is a function of random effects and fixed effects in both between- and within-level models. Using the multilevel CFA model, the total variance-covariance of observations may be expressed as a combination of three components in two levels: (a) factor loadings between indicators and latent factors ($\boldsymbol{\Lambda}_B$ and $\boldsymbol{\Lambda}_W$), (b) latent factor variances and covariance (explained portion of observed variance, $\boldsymbol{\Psi}_B$ and $\boldsymbol{\Psi}_W$), and (c) residual variance of indicators (unexplained portion of observed variance, $\boldsymbol{\Theta}_B$ and

$\boldsymbol{\Theta}_W$). The single-factor intraclass correlation (ICC) of MCFA is then defined as $\mathbf{ICC} = \boldsymbol{\Psi}_{\mathrm{B}}(\boldsymbol{\Psi}_{\mathbf{B}} + \boldsymbol{\Psi}_{\mathbf{W}})^{-1}$ (Muthén, 1991, 1994).

When the maximum modeling technique is applied to analyze these two-level data, Equation (4) becomes

$$Cov\left(\mathbf{y}_{ig}\right) = \boldsymbol{\Sigma}_B^{Saturated} + \boldsymbol{\Lambda}_W \boldsymbol{\Psi}_W \boldsymbol{\Lambda}_W{}' + \boldsymbol{\Theta}_W \quad \text{(Max CFA with saturated between -level structure, 5.1)}$$

or

$$Cov\left(\mathbf{y}_{ig}\right) = \boldsymbol{\Lambda}_B \boldsymbol{\Psi}_B \boldsymbol{\Lambda}_B{}' + \boldsymbol{\Theta}_B + \boldsymbol{\Sigma}_W^{Saturated} \quad \text{(Max CFA with saturated within -level structure, 5.2)}$$

The full-rank variance-covariance matrix $\boldsymbol{\Sigma}_B^{Saturated}$ or $\boldsymbol{\Sigma}_W^{Saturated}$ is unstructured, that is, all the possible between-level or within-level variation of indicators is estimated and separated from their total variance component. For the multilevel measurement model specification in this study, the unique within-level or between-level variation is then used to construct the respective within-level or between-level model with fixed and random effects without contamination from the other level. The residual part is the unique portion of total variation to the within-level or between-level variation of the indicators. If $\boldsymbol{\Lambda}_B = \boldsymbol{\Lambda}_W$ (i.e., equality of factor loadings and structures holds for between-/within-level models), the resulting factor loading estimates of design-based approach with one-level model are equal to the between-/within-level factor loadings in the true two-level model (i.e., $\boldsymbol{\Lambda}_B = \boldsymbol{\Lambda}_W = \boldsymbol{\Lambda}_{\mathbf{y}}$).

If we ignore the multilevel structure and construct a one-level model with design-based approach for the multilevel dataset $\mathbf{y}$, the observed variance-covariance matrix of the indicators may be represented with the model-driven parameters as follows:

$$Cov\left(\mathbf{y}\right) = \boldsymbol{\Lambda}_{\mathbf{y}} \boldsymbol{\Psi} \boldsymbol{\Lambda}_{\mathbf{y}}{}' + \boldsymbol{\Theta}_{\boldsymbol{\varepsilon}} = \boldsymbol{\Lambda}_{\mathbf{y}} \left(\boldsymbol{\Psi}_B + \boldsymbol{\Psi}_W\right) \boldsymbol{\Lambda}_{\mathbf{y}}{}' + \left(\boldsymbol{\Theta}_B + \boldsymbol{\Theta}_W\right) (1 - \text{level CFA}) \tag{6}$$

With the inclusion of ICC, Equation (4) can be further reformatted as:

$$Cov\left(\mathbf{y}\right) = \boldsymbol{\Lambda}_B \, \mathbf{ICC} \, \boldsymbol{\Psi} \boldsymbol{\Lambda}_B^{'} + \boldsymbol{\Lambda}_{\mathbf{W}} (\mathbf{I} - \mathbf{ICC}) \, \boldsymbol{\Psi} \boldsymbol{\Lambda}_{\mathbf{W}}' + \left(\boldsymbol{\Theta}_B + \boldsymbol{\Theta}_W\right) \tag{7}$$

However, if the magnitudes of non-zero elements in between-/within-level factor loading matrix are not the same, the factor loading estimates of design-based approach is a function of true between- and within-level factor loadings and the ICC measures. Snijders and Bosker (2011) shows that, in univariate case, the regression coefficient of overall model with multilevel dataset will be $\lambda_y = ICC\lambda_B + (1 - ICC)\,\lambda_W$. In the MCFA case, if there is a uni-factor structure in both levels, we hypothesize that the factor loading estimates of design-based approach could be simplified as (which is later being validated by the simulation result):

$$\boldsymbol{\Lambda}_y = ICC\boldsymbol{\Lambda}_B + (1 - ICC)\,\boldsymbol{\Lambda}_W \tag{8}$$

That is, design-based approach could yield a conflated factor loading estimate ($\boldsymbol{\Lambda}_y$) of complex survey data. If the indicator has more variation in the within level, its factor loading estimate from design-based approach will be close to its within-level counterpart; if the indicator has more variation in the between level, its conflated factor loading estimate will be close to its between-level counterpart.

The composite reliability with congeneric measures based on CFA can then be calculated for the above models (Raykov, 2004; Brown, 2006), using:

$$\rho = \left(\sum_{\mathrm{p}=1}^{\mathrm{P}} \lambda_{\mathrm{p}}\right)^2 \Bigg/ \left[\left(\sum_{\mathrm{p}=1}^{\mathrm{P}} \lambda_{\mathrm{p}}\right)^2 + \sum_{\mathrm{p}=1}^{\mathrm{P}} \Theta_{\mathrm{p}}\right], \tag{9}$$

where λ_p is the factor loading of item p onto a single common factor and Θ_p is the unique variance of item p. When constructing a one-level model, we can insert Equation (8) into (9) to obtain the reliability for the design-based model in Equation (10), which can be further expressed as the function of between- and within-level factor loadings and errors:

$$\rho_{Design-Based\ Approach} = \frac{\left[\left(\sum\limits_{p=1}^{P} ICC\lambda_{Bp}\right)^2 + \left(\sum\limits_{p=1}^{P} (1 - ICC)\,\lambda_{Wp}\right)^2\right]}{\left\{\left[\left(\sum\limits_{p=1}^{P} ICC\lambda_{Bp}\right)^2 + \left(\sum\limits_{p=1}^{P} (1 - ICC)\,\lambda_{Wp}\right)^2\right] + \sum\limits_{p=1}^{P} \left(\Theta_{Bp} + \Theta_{Wp}\right)\right\}}, \tag{10}$$

Where λ_{BP} and λ_{Wp} are the standardized factor loadings of item p in the between- and within-level, and Θ_{BP} and Θ_{Wp} are residual variances of item p in the between- and within-level. The detailed discussion about reliability measures in complex survey data with MCFA and CFA can be referred to Geldhof et al. (2013).

In the following sections, the simulation study was provided to illustrate the robustness of the three SEM modeling strategies with five model specifications in analyzing a measurement dataset obtained from complex survey. The simulation results could inform the influences of different modeling techniques on overall exact model fit chi-square test and various model fit indices, information criteria, parameter and standard error estimates as well as the statistical inferences in the statistical analysis. Parameter Specification for the Simulation From a substantive-methodological synergy (Marsh and Hau, 2007) perspective, we specified the population parameters in our simulations based on the parameter estimates obtained from an empirical dataset to examine the performance of the proposed modeling approaches on multilevel measurement data.

Empirical Dataset: Measurement and Sampling

From a sample of 784 academically at-risk children participating in a longitudinal study, we selected a balanced dataset of 120 students nested within 40 classrooms with 3 students in each class. A total of 120 students (47 Females and 73 males; 39 African Americans, 38 Hispanics, 40 Caucasians and 3 Asians/Pacific Islanders) were drawn. No evidence of selective consent for participation in the larger longitudinal study was found. Details about recruitment of multilevel sampling procedure of the 784 participants were reported in Hughes and Kwok (2007). Their Grade 3 Harter competence measures were used in the current study. We generated the balanced-design synthetic datasets based on the parameter estimates from the MCFA of their Harter competence measures, considering different levels of cluster sizes, cluster numbers and intraclass correlations.

The Children Perceived Competence Scale (CPC, Harter, 1982) is composed of three domain-specific competences, including child-perceived competence in scholastic competence (CPCSC), social acceptance (CPCSA), and athletic competence (CPCAC), as well as a general global self-worth scale (CPCSW). The item-level responses consisted of ordered and categorical 4-point scale. Each of the subscale was measured using 7 items for a total of 28 items. Reliability of the item-level subscales ranged from 0.75 to 0.86. We used the composite scores of each subscale to form four continuous indicators for children's general competence at both classroom and individual levels so that the analysis result can be generalized to continuous responses.

Simulation Study: True Model Specification

In order to demonstrate the adequacy and robustness of five different modeling approaches, we used Monte-Carlo simulation to generate the synthetic complex survey dataset with known true multilevel measurement model of CPC scale. A two-level uni-factor CFA model was firstly built for the empirical dataset of CPC scale with an overall factor of child-perceived competence including three domain-specific subscale indicators and one general self-worth indicator in both the between- and within-levels for the empirical dataset (as shown in **Figure 1**). With Full Information Maximum Likelihood (FIML) estimation, the resulting two-level CFA has an adequate model fit test statistic and index values ($\chi^2 = 9.421$ with $df = 4$ and $p = 0.051$, $CFI = 0.990$, $RMSEA = 0.048$, $SRMR\text{-}Within = 0.023$, $SRMR\text{-}Between = 0.018$). The parameter estimates of varying factor loadings were retained in the true models for simulation. The ICCs for the indicators in the empirical analysis ranged from 0.352 to 0.617. The factor variances in between- and within-level would then be altered to have different ICC settings in the simulation study.

Even though the between- and within-level model had equal structures, their factor loading magnitudes and patterns of significance were distinct for this empirical dataset (see the two dashed lines in **Figure 1B**). The unstandardized factor loading estimates from the two-level CFA analysis of empirical dataset were used as the population values for Monte Carlo simulation. The population values for the within-level factor loadings was 1 for scholastic competence (marker variable with standardized factor loading $\lambda = 0.719$), 0.45 for social acceptance ($\lambda = 0.400$), 0.92 for athletic competence ($\lambda = 0.694$), and 0.36 for global self-worth ($\lambda = 0.331$). In the within-level, only athletic competence was a statistically significant factor loading (i.e., $p \leq 0.05$). On the other hand, all the between-level factor loadings were statistically significant. The between-level factor loadings were 1 (marker variable with standardized factor loading $\lambda = 0.910$) for scholastic competence, 0.78 for social acceptance ($\lambda = 0.871$), 0.60 for athletic competence ($\lambda = 0.807$), and 0.62 for global self-worth ($\lambda = 0.816$). The intercepts of the indicators were set as 2.896 for scholastic competence, 2.856 for social acceptance, 2.860 for athletic competence, and 3.268 for global self-worth. Finally, the population values of residual variance for the classroom- and individual-level indicators were set as 0.2 and 0.5. Total variance of factor was set at one ($\Psi_{CPC} = \Psi_{B\_CPC} + \Psi_{W\_CPC} = 1$), and the between- and within-level factor variance was set as Ψ_{B_CPC} and Ψ_{W_CPC}. The levels of intraclass correlation (ICC) were then manipulated as $\Psi_{B\_CPC}(\Psi_{B\_CPC} + \Psi_{W\_CPC})^{-1}$. For the simulation study, the true two-level model was constructed with these empirical parameter estimates under varying conditions of cluster size (CS = 3, 30, 200), cluster

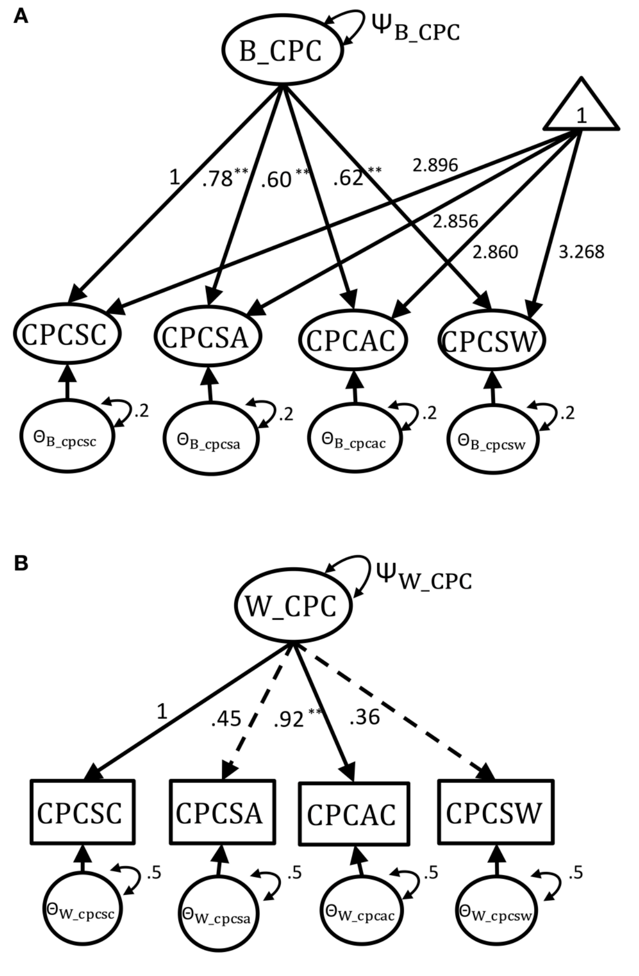


FIGURE 1 | The multilevel CFA model with parameters from empirical Harter dataset. **(A)** The true between-level model. **(B)** The true within-level model. $^{**}p < 0.05$.

number (CN = 40, 100, 300) and Intraclass correlation (ICC = 0.1, 0.3, 0.5, 0.7, 0.9, Muthén, 1994) to generate 1,000 converged copies of balanced-design complex survey datasets. A total of 3(CS)* 3(CN)*5(ICC)*1,000(reps) = 45,000 synthetic multilevel datasets were generated.

Simulation Study: Analytical Models Specification

Five SEM models for multilevel data with robust estimation were used to analyze the synthetic datasets. For ease of differentiation, we used the following naming scheme for the five model specifications:

(1) 2MLR: the two-level model-based model and the true model (**Figures 1A,B**).
(2) 1MLR: the one-level design-based model (**Figure 1B**).
(3) 2MaxB[1]: the two-level maximum model with saturated model in between level (**Figure 2**) and true model in within level (**Figure 1B**).
(4) 2MaxW: the two-level maximum model with true model in between level (**Figure 1A**) and saturated model in within level (**Figure 2**).
(5) 2Miss: the miss-specified two-level model was constructed as **Figures 1A,B** by constraining the factor loading estimates of the between and within levels to be the same. This miss-specified model was used to test if the model-based approach is robust in detecting trivial items, and to examine if this model performs the same as design-based approach (i.e., 1MLR).

Two Mplus built-in routines were employed for the statistical modeling (Muthén and Muthén, 2012). First, the TYPE = TWOLEVEL routine, which allows level-specific specifications for complex survey data, was used for the 2MLR, 2MaxB, 2MaxW, and 2Miss). Second, TYPE = COMPLEX was used as design-based approach, where only a single level model is estimated (i.e., 1MLR) for complex survey data. By default, both routines use the full information maximum likelihood (FIML) parameter estimator and the robust standard error estimator; in Mplus, this procedure is called as maximum likelihood estimation with robust standard error correction (MLR), which is useful for non-normal and non-independent observations (Muthén and Satorra, 1995). Different from using the inverse of information matrix as the sampling variance estimate with normal distribution assumption, an asymptotically consistent estimate of covariance matrix is derived directly from observations by including a scaling matrix in between two copies of the Hessian matrix and then is used to compute the robust estimate of sampling variance, which is the square of standard error (Huber, 1967; White, 1980; Hardin and Hilbe, 2007). The chi-square test statistic reported using MLR is asymptotically equivalent to Yuan-Bentler T2* test statistic (Muthén and Muthén, 2012). We compared each model performance in simulation convergence rate (CR), model-fit test statistic and fit indices, Information Criteria (i.e., AIC, BIC and adjusted BIC), and the estimates of between/within-level factor loadings, scale reliabilities, residual variance and mean structure estimates as well as their 95% coverage rate and empirical power. Level-specific scale reliability was calculated based on Geldhof et al. (2013) using Equation (9) to decompose variance in an item into the individual component and the cluster component.

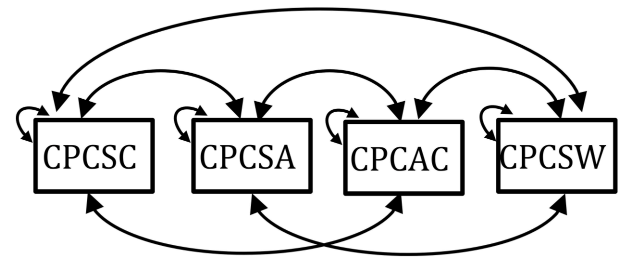


FIGURE 2 | The saturated model.

RESULTS

Convergence Rate of Simulations, Model Fit Test Statistic, Fit Indices and Information Criteria

For ease of illustration, we selected the results of simulation conditions with the smallest cluster number (CN = 40 with CS = 3, 30, 200) and the largest combination of sample size (CN = 300 with CS = 200) in **Figure 3**. When CN larger than 40, the five modeling techniques achieved convergent results across different ICC conditions. Nevertheless, with a cluster of 40, the convergence ratio varied with ICC values: 2Miss and 1MLR reached 100% convergence for all ICCs, but 2MLR, 2MaxB and 2MaxW had 9.5∼38.2% non-convergent simulation results when ICC was smaller than 0.3 or larger than 0.7. For instance, in the smallest case of CN(CS) = 40(3), the CR pattern of the five modeling techniques differed with ICC values: 1MLR and 2Miss reached perfect convergence in all ICC conditions; the CR for 2MaxW exhibited a quadratic pattern, which increased with the increase of ICC and leveled off and reached 100% when ICC ≥0.5 while 2MaxB demonstrated a reversed pattern. 2MLR had a downward-U quadratic pattern of CRs verse ICCs with the peak at ICC = 0.5. According to the error message, the non-convergent result of 2-level models was mostly due to the non-positive definite first-order derivative product matrix for the insufficient portion of variance in the within or between level, especially in the smaller sample size conditions.

All models yielded significant Chi-square exact test results but adequate CFIs, RMSEAs, and SRMR-W values (e.g., *CFI* > 0.90, *RMSEA* < 0.08 and *SRMR* < 0.08, Hu and Bentler, 1999) in all simulation conditions. However, for SRMR-Bs, 2Miss consistently demonstrated badness of fit across most of simulation conditions. Particularly, the SRMS-Bs of the 2Miss showed a quadratic pattern with downward-U shape and peaked between 0.5 and 0.7 for models with a sample size equal or greater than CN(CS) = 40(30). The result suggested that the

[1] The exemplary Mplus syntax of 2MaxB model is provided in **Appendix** for reader's reference.

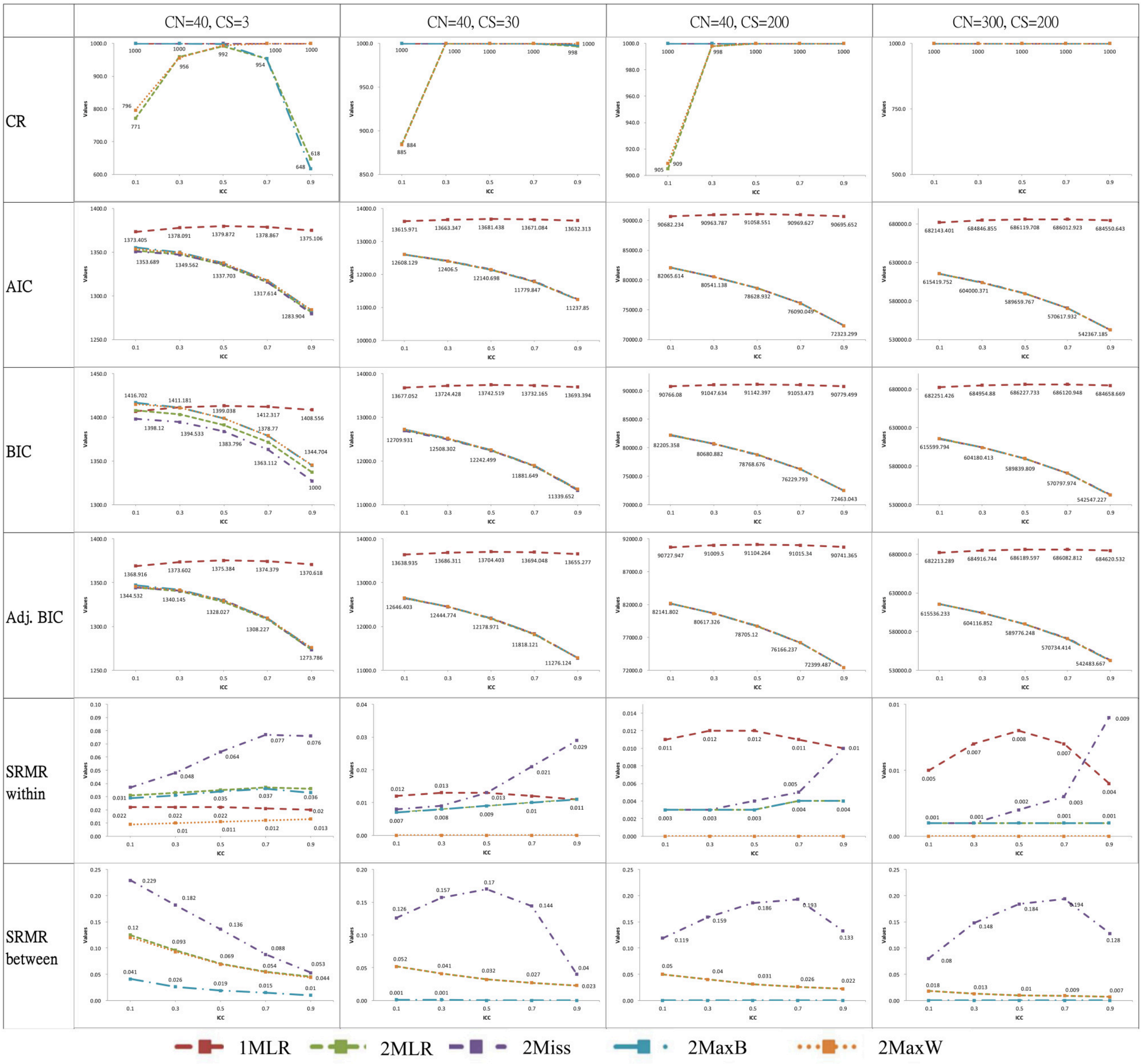


FIGURE 3 | Plots of selected analytical outputs of ICC against fit statistics across different modeling strategies. CN, Cluster number; CS, Cluster size; CR, Convergence rate of simulations; ICC, Intraclass correlation. 1MLR, the one-level design-based model; 2MLR, the two-level model-based model and the true model; 2MaxB, the two-level maximum model with saturated model in between level and true model in within level; 2MaxW, the two-level maximum model with true model in between level and saturated model in within level; 2Miss, the miss-specified two-level model by constraining the factor loading estimates of the between and within levels to be the same.

2Miss showed lack of fit to the multilevel measurement dataset with level-varying parameters.

Across all simulation conditions, four 2-level models consistently generated smaller AIC and adj. BIC than the 1MLR. The average difference of AIC and adj. BIC between 2-level models and 1MLR were larger than 20 for all the simulation cases even for the smallest sample size conditions (e.g., for [CN(CS), ICC] = [40(3), 0.1], $\text{AIC}_{\text{1MLR}} = 1{,}373.41$ vs. $\text{AIC}_{\text{2MaxB}} = 1{,}355.38$, and adj. $\text{BIC}_{\text{1MLR}} = 1{,}368.92$ vs. adj. $\text{BIC}_{\text{2MaxB}} = 1{,}347.15$). AIC and adj. BIC indices preferred model-based approaches over design-based approaches across all simulation settings. BIC could distinguish the 2-level models from 1MLR in most of simulation conditions, but not for the conditions with the smallest sample CN(CS) = 40(3) at ICC < 0.3.

Estimation of Fixed Effects

The parameter estimates of [CN(CS), ICC] = [300(200), 0.3], [40(30), 0.3] and [40(3), 0.3] were summarized in **Tables 1–3**. Besides, the relative and absolute bias values of estimated factor loadings of CPCSA and CPCAC were tabulated in **Table 4**[2].

[2] Because CPCSC is the maker variable so its factor loading would constantly fixed at one for all the analytical models. Therefore, we didn't present its bias measures.

TABLE 1 | Simulated unstandardized results[a] of SEM techniques on synthetic harter's competence dataset for [CN(CS), ICC] = [300(200), 0.3].

	Population settings	2MLR	1MLR	2MaxB	2Miss	2MaxW
Convergence rate	----	1.00	1.00	1.00	1.00	1.00
Chi-square (df)	----	4.113 (4)	6.490 (2)	2.057(2)	80.343 (7)	2.082 (2)
CFI	----	1.00	0.998	1.00	0.998	1.00
RMSEA	----	0.001	0.005	0.001	0.013	0.001
SRMR(SRMR_B)	---- (----)	0.001 (0.013)	0.007	0.001(<0.001)	0.001 (0.148)	<0.001 (0.013)
AIC	----	604,009.427	684,769.911	604,011.412	604,079.363	604,011.391
BIC	----	604,189.469	684,877.936	604,209.458	604,232.398	604,209.437
ABIC	----	604,125.909	684,839.800	604,139.542	604,178.372	604,139.520

	Population		Est.	SE	95%	Sig.	Est.	SE	95%	Sig	Est.	SE	95%	Sig	Est.	SE	95%	Sig	Est.	SE	95%	Sig
WITHIN-LEVEL																						
W_CPC by																						
CPCSC	1.000		1.000	–	–	–	1.000	–	–	–	1.000	–	–	–	1.000	–	–	–				
CPCSA	0.450		0.450	0.005	0.944	1.00	0.534	0.021	0.023	1.00	0.450	0.005	0.944	1.00	0.451	0.005	0.000	1.00				
CPCAC	0.920	**	0.920	0.008	0.945	1.00	0.791	0.033	0.036	1.00	0.920	0.008	0.945	1.00	0.918	0.008	0.000	1.00				
CPCSW	0.360		0.360	0.005	0.953	1.00	0.428	0.020	0.061	1.00	0.360	0.005	0.953	1.00	0.361	0.005	0.000	1.00				
Ψ_{W_CPC}	0.700	**	0.700	0.009	0.955	1.00	1.036	0.052	0.000	1.00	0.700	0.009	0.955	1.00	0.701	0.008	0.959	1.00				
Residual Variance																						
CPCSC	0.500	**	0.500	0.006	0.951	1.00	0.662	0.035	0.003	1.00	0.500	0.006	0.951	1.00	0.499	0.006	0.947	1.00				
CPCSA	0.500	**	0.500	0.003	0.943	1.00	0.729	0.021	0.000	1.00	0.500	0.003	0.943	1.00	0.500	0.003	0.943	1.00				
CPCAC	0.500	**	0.500	0.006	0.951	1.00	0.753	0.028	0.000	1.00	0.500	0.006	0.951	1.00	0.501	0.006	0.947	1.00				
CPCSW	0.500	**	0.500	0.003	0.948	1.00	0.716	0.019	0.000	1.00	0.500	0.003	0.948	1.00	0.500	0.003	0.950	1.00				
BETWEEN-LEVEL																						
B_CPC by																						
CPCSC	1.000		1.000	–	–	–									1.000	–	–	–	1.000	–	–	–
CPCSA	0.780	**	0.785	0.084	0.949	1.00									0.451	0.005	0.000	1.00	0.785	0.084	0.949	1.00
CPCAC	0.600	**	0.603	0.071	0.955	1.00									0.918	0.008	0.000	1.00	0.603	0.071	0.955	1.00
CPCSW	0.620	**	0.622	0.073	0.949	1.00									0.361	0.005	0.000	1.00	0.622	0.073	0.949	1.00
Ψ_{B_CPC}	0.300		0.301	0.045	0.941	1.00									0.255	0.029	0.630	1.00	0.301	0.045	0.941	1.00
Residual variance																						
CPCSC	0.200	**	0.197	0.030	0.939	0.999									0.239	0.031	0.777	1.00	0.197	0.030	0.939	1.00
CPCSA	0.200	**	0.199	0.023	0.945	1.00									0.273	0.024	0.102	1.00	0.199	0.023	0.945	1.00
CPCAC	0.200	**	0.198	0.020	0.935	1.00									0.153	0.022	0.415	1.00	0.198	0.020	0.935	1.00
CPCSW	0.200	**	0.199	0.020	0.937	1.00									0.244	0.021	0.459	1.00	0.199	0.020	0.937	1.00
INTERCEPT/MEAN																						
CPCSC	2.896	**	2.898	0.041	0.937	1.00	2.898	0.041	0.937	1.00	2.898	0.041	0.937	1.00	2.898	0.041	0.937	1.00	2.898	0.041	0.937	1.00
CPCSA	2.856	**	2.857	0.036	0.957	1.00	2.857	0.036	0.957	1.00	2.857	0.036	0.957	1.00	2.857	0.036	0.957	1.00	2.857	0.036	0.957	1.00
CPCAC	2.860	**	2.862	0.032	0.943	1.00	2.862	0.032	0.943	1.00	2.862	0.032	0.043	1.00	2.862	0.032	0.943	1.00	2.862	0.032	0.943	1.00
CPCSW	3.268	**	3.268	0.033	0.946	1.00	3.268	0.033	0.946	1.00	3.268	0.033	0.946	1.00	3.268	0.033	0.946	1.00	3.268	0.033	0.946	1.00

*Ψ_{B_CPC} and Ψ_{W_CPC} is the between-/within-level factor variance. The normal font indicates the fixed effect and intercept estimate; the italic indicates the random effect estimate. Est, estimate; SE, standard error; 95%, 95% confidence interval coverage rate; Sig., empirical power. **$p < 0.05$.*

[a] *The standardized result can be requested from the author.*

TABLE 2 | Simulated unstandardized results[a] of SEM techniques on synthetic harter's competence dataset for [CN(CS), ICC] = [40(30), 0.3].

	Population settings		2MLR				1MLR				2MaxB				2Miss				2MaxW			
Convergence rate	----		0.999				1.00				1.00				1.00				0.998			
Chi-square (df)	----		4.935 (4)				3.041 (2)				2.041 (2)				16.832 (7)				3.280 (2)			
CFI	----		0.998				0.995				0.999				0.988				0.998			
RMSEA	----		0.011				0.015				0.010				0.015				0.015			
SRMR(SRMR_B)	---- (----)		0.008 (0.043)				0.013				0.008 (0.001)				0.009 (0.158)				<0.001 (0.043)			
AIC	----		12,406.494				13,661.328				12,408.438				12,412.245				12,408.726			
BIC	----		12,508.295				13,722.409				12,520.42				12,498.776				12,520.708			
ABIC	----		12,444.768				13,684.292				12,450.54				12,444.777				12,450.828			
	Population		Est.	SE	95%	Sig.	Est.	SE	95%	Sig	Est.	SE	95%	Sig	Est.	SE	95%	Sig	Est.	SE	95%	Sig
WITHIN-LEVEL																						
W_CPC by																						
CPCSC	1.000		1.000	–	–	–	1.000	–	–	–	1.000	–	–	–	1.000	–	–	–				
CPCSA	0.450		0.451	0.034	0.936	1.00	0.533	0.064	0.757	1.00	0.450	0.034	0.937	1.00	0.457	0.034	0.937	1.00				
CPCAC	0.920	**	0.922	0.059	0.929	1.00	0.795	0.097	0.680	1.00	0.922	0.059	0.929	1.00	0.907	0.058	0.918	1.00				
CPCSW	0.360		0.359	0.032	0.949	1.00	0.426	0.060	0.816	1.00	0.359	0.032	0.950	1.00	0.365	0.032	0.948	1.00				
Ψ_{W_CPC}	0.700	**	0.700	0.602	0.932	1.00	1.041	0.156	0.421	1.00	0.703	0.060	0.932	1.00	0.708	0.060	0.931	1.00				
Residual Variance																						
CPCSC	0.500	**	0.500	0.045	0.941	1.00	0.653	0.107	0.679	1.00	0.498	0.045	0.941	1.00	0.494	0.045	0.933	1.00				
CPCSA	0.500	**	0.500	0.022	0.943	1.00	0.719	0.061	0.021	1.00	0.499	0.022	0.943	1.00	0.497	0.023	0.936	1.00				
CPCAC	0.500	**	0.500	0.040	0.943	1.00	0.743	0.083	0.156	1.00	0.498	0.040	0.942	1.00	0.507	0.039	0.931	1.00				
CPCSW	0.500	**	0.500	0.022	0.937	1.00	0.708	0.057	0.009	1.00	0.499	0.022	0.937	1.00	0.498	0.022	0.935	1.00				
BETWEEN-LEVEL																						
B_CPC by																						
CPCSC	1.000		1.000	–	–	–									1.000	–	–	–	1.000	–	–	–
CPCSA	0.780	**	1.042	0.421	0.925	0.874									0.457	0.034	0.000	1.00	0.928	0.542	0.926	0.876
CPCAC	0.600	**	0.621	0.244	0.933	0.801									0.907	0.058	0.000	1.00	0.621	0.245	0.935	0.801
CPCSW	0.620	**	0.653	0.262	0.928	0.819									0.365	0.032	0.000	1.00	0.651	0.262	0.929	0.820
Ψ_{B_CPC}	0.300		0.348	0.494	0.905	0.664									0.261	0.091	0.817	0.935	0.314	0.147	0.907	0.665
Residual Variance																						
CPCSC	0.200	**	0.145	0.455	0.935	0.598									0.230	0.088	0.939	1.00	0.179	0.108	0.937	0.597
CPCSA	0.200	**	0.163	0.102	0.907	0.799									0.261	0.064	0.893	1.00	0.172	0.120	0.906	0.801
CPCAC	0.200	**	0.186	0.056	0.889	0.950									0.154	0.064	0.773	1.00	0.187	0.056	0.888	0.950
CPCSW	0.200	**	0.186	0.058	0.889	0.942									0.236	0.057	0.927	1.00	0.186	0.058	0.890	0.940
INTERCEPT/MEAN																						
CPCSC	2.896	**	2.896	0.115	0.945	1.00	2.896	0.116	0.947	1.00	2.896	0.115	0.945	1.00	2.896	0.115	0.945	1.00	2.896	0.115	0.945	1.00
CPCSA	2.856	**	2.852	0.099	0.936	1.00	2.852	0.100	0.939	1.00	2.852	0.099	0.936	1.00	2.852	0.099	0.936	1.00	2.852	0.099	0.936	1.00
CPCAC	2.860	**	2.862	0.091	0.933	1.00	2.872	0.092	0.935	1.00	2.862	0.091	0.933	1.00	2.862	0.091	0.933	1.00	2.862	0.091	0.934	1.00
CPCSW	3.268	**	3.270	0.090	0.947	1.00	3.270	0.091	0.948	1.00	3.270	0.090	0.947	1.00	3.270	0.090	0.947	1.00	3.270	0.090	0.947	1.00

*Ψ_{B_CPC} and Ψ_{W_CPC} is the between-/within-level factor variance. The normal font indicates the fixed effect and intercept estimate; the italic indicates the random effect estimate. Est, estimate; SE, standard error; 95%, 95% confidence interval coverage rate; Sig., empirical power. **$p < 0.05$.*

[a] *The standardized result can be requested from the author.*

TABLE 3 | Simulated unstandardized results[a] of SEM techniques on synthetic harter's competence dataset for [CN(CS), ICC] = [40(3), 0.3].

	Population settings		2MLR				1MLR				2MaxB				2Miss				2MaxW			
Convergence rate	----		0.786				1.000				0.820				0.879				0.782			
Chi-square (df)	----		8.674 (4)				2.680 (2)				27.99 (2)				12.63 (7)				8.111 (2)			
CFI	----		0.959				0.982				0.968				0.941				0.963			
RMSEA	----		0.056				0.042				0.078				0.063				0.070			
SRMR(SRMR_B)	---- (----)		0.033 (0.085)				0.023				0.032 (0.025)				0.050 (0.173)				0.010 (0.082)			
AIC	----		1,347.312				1,368.669				1,347.010				1,347.254				1,349.315			
BIC	----		1,403.062				1,402.119				1,408.335				1,394.641				1,410.640			
ABIC	----		1,339.831				1,364.181				1,338.781				1,340.895				1,341.086			
	Population		Est.	SE	95%	Sig.	Est.	SE	95%	Sig	Est.	SE	95%	Sig	Est.	SE	95%	Sig	Est.	SE	95%	Sig
WITHIN-LEVEL																						
W_CPC by																						
CPCSC	1.000		1.000	–	–	–	1.000	–	–	–	1.000	–	–	–	1.000	–	–	–				
CPCSA	0.450		0.453	0.146	0.944	0.808	0.607	0.154	0.891	0.988	0.474	0.147	0.954	0.829	0.509	0.123	0.941	0.991				
CPCAC	0.920	**	0.952	0.287	0.949	0.951	0.910	0.215	0.965	0.998	0.988	0.286	0.980	0.965	0.839	0.179	0.863	0.996				
CPCSW	0.360		0.358	0.137	0.944	0.796	0.484	0.140	0.909	0.963	0.374	0.140	0.951	0.819	0.407	0.112	0.939	0.960				
Ψ_{W_CPC}	0.700	**	0.723	0.265	0.938	0.881	1.061	0.246	0.994	0.992	0.654	0.232	0.935	0.904	0.720	0.214	0.925	0.973				
Residual Variance																						
CPCSC	0.500	**	0.486	0.206	0.964	0.732	0.739	0.191	0.781	0.983	0.514	0.183	0.962	0.807	0.479	0.165	0.937	0.829				
CPCSA	0.500	**	0.491	0.085	0.924	0.999	0.693	0.115	0.623	1.00	0.488	0.085	0.919	1.00	0.485	0.086	0.915	1.00				
CPCAC	0.500	**	0.470	0.174	0.947	0.796	0.683	0.165	0.755	0.967	0.456	0.169	0.953	0.787	0.542	0.139	0.927	0.952				
CPCSW	0.500	**	0.487	0.081	0.922	1.00	0.687	0.105	0.588	1.00	0.484	0.081	0.913	0.999	0.484	0.081	0.916	1.00				
BETWEEN-LEVEL																						
B_CPC by																						
CPCSC	1.000		1.000	–	–	–									1.000	–	–	–	1.000	–	–	–
CPCSA	0.780	**	0.930	0.812	0.926	0.409									0.509	0.123	0.356	0.991	1.128	0.830	0.908	0.419
CPCAC	0.600	**	0.681	0.582	0.964	0.428									0.839	0.179	0.784	0.996	0.667	0.536	0.965	0.438
CPCSW	0.620	**	0.721	0.747	0.945	0.363									0.407	0.112	0.457	0.960	0.700	0.594	0.941	0.375
Ψ_{B_CPC}	0.300		0.337	0.289	0.935	0.200									0.338	0.198	0.950	0.340	0.342	0.288	0.934	0.195
Residual Variance																						
CPCSC	0.200	**	0.158	0.210	0.955	0.183									0.165	0.146	0.909	0.173	0.156	0.214	0.955	0.190
CPCSA	0.200	**	0.140	0.189	0.938	0.302									0.233	0.101	0.930	0.647	0.109	0.213	0.934	0.299
CPCAC	0.200	**	0.162	0.140	0.930	0.268									0.164	0.116	0.891	0.251	0.164	0.135	0.932	0.272
CPCSW	0.200	**	0.154	0.166	0.932	0.354									0.215	0.092	0.934	0.678	0.158	0.137	0.928	0.364
INTERCEPT/MEAN																						
CPCSC	2.896	**	2.903	0.149	0.943	1.00	2.909	0.140	0.932	1.00	2.903	0.147	0.933	1.00	2.903	0.149	0.939	1.00	2.902	0.149	0.941	1.00
CPCSA	2.856	**	2.862	0.121	0.938	1.00	2.863	0.120	0.957	1.00	2.861	0.120	0.948	1.00	2.862	0.120	0.942	1.00	2.862	0.121	0.938	1.00
CPCAC	2.860	**	2.863	0.129	0.946	1.00	2.867	0.127	0.948	1.00	2.863	0.127	0.942	1.00	2.864	0.129	0.944	1.00	2.863	0.129	0.943	1.00
CPCSW	3.268	**	3.264	0.112	0.941	1.00	3.267	0.112	0.938	1.00	3.267	0.111	0.934	1.00	3.268	0.111	0.936	1.00	3.264	0.112	0.938	1.00

*Ψ_{B_CPC} and Ψ_{W_CPC} is the between-/within-level factor variance. The normal font indicates the fixed effect and intercept estimate; the italic indicates the random effect estimate. Est, estimate; SE, standard error; 95%, 95% confidence interval coverage rate; Sig., empirical power. **$p < 0.05$.*

[a] *The standardized result can be requested from the author.*

TABLE 4 | The relative bias and absolute bias of factor loading estimates from five SEM modeling techniques for ICC = 0.3.

CN(CS)	Model	Within Level				Between Level			
		CPCSA		CPCAC		CPCSA		CPCAC	
		Bias (%)	Abs(Bias) (%)	Bias (%)	Abs(Bias) (%)	Bias (%)	Abs(Bias) (%)	Bias (%)	Abs(Bias) (%)
40(3)	2MLR	0.72	24.58	3.52	21.88	19.18	53.69	13.41	52.32
	2MaxB	5.21	24.24	7.35	20.57				
	2MaxW					19.55	54.57	11.11	49.29
	2Miss	13.13	23.66	−8.84	17.47	−34.74	35.32	39.78	41.25
	1MLR	20.79	28.16	−12.02	19.03				
	1MLR*	−0.99	19.13	−1.77	17.03				
40(30)	2MLR	0.23	6.17	0.19	5.27	14.63	39.07	3.55	30.88
	2MaxB	0.22	6.17	0.19	5.27				
	2MaxW					15.48	39.94	3.56	30.87
	2Miss	1.59	6.23	−1.43	5.37	−41.39	41.39	51.14	51.14
	1MLR	18.49	19.74	−13.55	14.92				
	1MLR*	2.48	12.03	−2.03	7.75				
300(200)	2MLR	−0.02	0.89	0.02	0.72	0.62	8.76	0.57	9.41
	2MaxB	−0.02	0.89	0.02	0.72				
	2MaxW					0.62	8.76	0.57	9.41
	2Miss	0.19	0.90	−0.22	0.75	−42.20	42.20	52.99	52.99
	1MLR	18.67	18.67	−14.01	14.01				
	1MLR*	−2.73	3.87	−3.92	4.63				

Bias: relative bias = $\frac{Estimate-Parameter}{Parameter}$ *; Abs(Bias) =* $\left|\frac{Estimate\ -Parameter}{Parameter}\right|$*. The parameter value of 2-level models and 1MLR in the within level:* $\lambda_{CPCSA} = 0.45$, $\lambda_{CPCAC} = 0.92$ *and in the between level:* $\lambda_{CPCSA} = 0.78$, $\lambda_{CPCAC} = 0.60$ *of population two-level model. 1MLR* presents the bias measures with respect to its true conflated parameter value from Equation (8):* $\lambda_{CPCSA} = 0.549$, $\lambda_{CPCAC} = 0.824$*.*

CPCSC was the maker variable so its factor loading would constantly be fixed at one for all the analytical models. CPCSW and CPCSA had the same pattern of bias; therefore, we presented the result of for CPCSA and CPCAC only. Relative bias (RB) is calculated as the value of parameter estimate minus the population value divided by the population value. RB quantifies the degree of deviation of the parameter estimate relative to the population value. A zero value of RB reflects an unbiased estimate of the parameter. A negative value indicates an underestimation of the parameter; on the other hand, a positive value indicates an overestimation of the parameter. According to Flora and Curran (2004), the value of RB less than 5% is considered as trivial, between 5 and 10% as moderate, and greater than 10% as substantial. Absolute bias (AB) is the absolute value of RB, which will always be positive and cumulated to reflect the total amount of bias. Across simulation settings, 2MLR, 2MaxB, and 2MaxW models tended to generate factor loading estimates consistent with the population values in respective levels. The empirical results were consistent with the mathematical derivations [e.g., Equation (4), (5.1) and (5.2)]. Generally, as shown in **Table 4**, ABs were larger than their RB counterparts in smaller CN and CS, but as CN and CS increased, the discrepancy between RB and AB were smaller. The RBs and ABs of the parameter estimates were also getting smaller when sample size increased for 2-level models, except that 2Miss consistently generated biased between-level loading estimates across all sample size settings.

On the other hand, 1MLR and 2Miss tended to generate conflated estimates for the factor loadings, consistent with Equation (7). Take the condition of the smallest sample size as example [CN(CS), ICC] = [40(3), 0.3], compared with the within-level fixed effects in the population model, substantial relative bias was found in the factor loading estimates of 1MLR and 2Miss ranging from −12.02 to 20.79%. In contrast, negligible relative bias of factor loading estimates was found in the 2MLR and 2MaxB models ranging from 0.72 to 7.35% (e.g., $\lambda^{True\ model}_{CPCSA,\ W\_CPC} = 0.450$, $\hat{\lambda}^{2MLR}_{CPCSA,W\_CPC} = 0.453$ and $\hat{\lambda}^{2MaxB}_{CPCSA,W\_CPC} = 0.474$ vs. $\hat{\lambda}^{1MLR}_{CPCSA,CPC} = 0.607$, and $\hat{\lambda}^{2Miss}_{CPCSA,W\_CPC} = 0.509$). We also compared the factor loading estimates of 1MLR with its theoretical conflated values (obtained from Equation (8) with ICC = 0.3, e.g., $\lambda^{1MLR}_{CPCSA,CPC} = 0.549$) and presented the biases in **Table 4** at the row of 1MLR*. 1MLR generated negligible biases which grew larger as sample size increased (e.g., the relative bias ranged from −0.99 to −3.92%). Compared with the between-level fixed effects in the population model, the 2MLR and 2MaxW models yielded considerable relative and absolute biases at CN = 40 (the relative bias ranged

The bias measures for CPCSW had the same pattern with CPCSA; thus, we did not present the bias statistics, either.

from 3.55 to 19.55%; the absolute bias ranged from 54.57 to 30.87%).

To further investigate the relationship between factor loading estimates and sample sizes (e.g., CN×CS), we tabulated the between- and within-level $\hat{\lambda}$ of CPCSA and CPCAC in boxplots for ICC = 0.3 in **Figure 4**. The dispersion of the parameter estimates of the five models decreased as the sample size increased. When sample size was small, the dispersion of 2MLR was larger than 2MaxB/2MaxW. Across all cluster number and cluster size combinations, the 2MLR and the 2MaxB/2MaxW had consistent median estimates to their parameters. However, the 1MLR models generated conflated parameter estimates which would regress to the weighted means of the true factor loadings from the between-and within-level models (The true value of within-level $\lambda^{True\ model}_{CPCAC,\ W\_CPC} = 0.920$, between-level $\lambda^{True\ model}_{CPCAC,\ B\_CPC} = 0.600$, and the conflated parameter $\lambda^{1MLR}_{CPCSA,CPC} = 0.824$, vs. the estimate of $\hat{\lambda}^{1MLR}_{CPCAC,\ CPC} = 0.791$; $\lambda^{True\ model}_{CPCSA,\ W\_CPC} = 0.450$, $\lambda^{True\ model}_{CPCSA,\ B\_CPC} = 0.780$, and $\lambda^{1MLR}_{CPCSA,CPC} = 0.549$ vs. $\hat{\lambda}^{1MLR}_{CPCSA,\ CPC} = 0.534$). Different from 1MLR, the 2Miss models had consistent and efficient factor loading estimates as those produced by the 2MLR and 2MaxB models when sample size was greater than 1,200 [i.e., CN(CS) = 40(30)] in the within-level models; whereas, the 2Miss models generated biased parameter estimates across all sample size conditions in the between-level level.

The Conflated Factor Loading Estimates in Design-Based Models As ICC Changes

To probe into the consequence of applying design-based approach on complex survey data, we plotted the estimates (solid lines) of factor loadings from simulations and those (dash lines) from mathematical derivation (see Equation 8) against different ICC values in **Figure 5**. As we expected from the mathematical derivation, the factor loading estimates of the design-based model approached the true between-level values as ICCs increased. Even though they were supposed to reflect the within-level information, the estimates got conflated across all simulated ICCs, except for ICC = 0.

Estimation of Random Effects

In terms of factor variance, the four 2-level models yielded consistent random effect estimates (e.g., for [40(3), 0.3], in the between level: $\hat{\Psi}^{2MLR}_{B\_CPC} = 0.337$, $\hat{\Psi}^{2MaxW}_{B\_CPC} = 0.342$ and $\hat{\Psi}^{2Miss}_{B\_CPC} = 0.338$; in the within level: $\hat{\Psi}^{2MLR}_{W\_CPC} = 0.723$, $\hat{\Psi}^{2MaxB}_{W\_CPC} = 0.654$, $\hat{\Psi}^{2Miss}_{W\_CPC} = 0.720$). The performance of the 1MLR, however, was not as consistent as that of the three 2-level models in estimating the random effects. Specifically, the factor variance estimate of 1MLR equaled 1.061, which was roughly the sum of the population between- and within-level factor variance values as shown in Equation (6). The substantial relative bias reached 51.57%. The 1MLR also yielded the same overall estimates for the residual variances (i.e., residuals of Equation 7), while the three 2-level models had fair within-level residual variance estimate (e.g., $\hat{\theta}^{1MLR}_{CPCSW} = 0.687$ vs. $\theta^{True\ model}_{CPCSW,\ Within-level} = 0.500$, $\hat{\theta}^{2MLR}_{CPCSW,\ Within-level} = 0.487$, $\hat{\theta}^{2MaxB}_{CPCSW,\ Within-level} = 0.484$ and $\hat{\theta}^{2Miss}_{CPCSW,\ Within-level} = 0.484$).

Mean Structures

As for the mean structure, all examined models yielded consistent mean/intercept estimates with conformable statistical inferences as shown in **Tables 1–3**.

The 95% Confidence Interval Coverage Rate and Empirical Power of Estimates

With the conflated parameter estimate of fixed and random effect, the 95% confidence interval coverage rate[3] (95%) of 1MLR and 2Miss tended to be much smaller than its nominal level. In terms of empirical power[4] (Sig.), all the empirical power for the three factor loading estimates were equal to or close to 1 in the 1MLR and 2Miss (e.g., for [40(3), 0.3], $\hat{\lambda}^{1MLR}_{CPCSA,CPC} = 0.659$, 95% = 0.891, Sig = 0.988 in **Table 3**). In contrast, in the 2MLR and 2MaxB models, the empirical power of $\hat{\lambda}_{CPCSA,W\_CPC}$ and $\hat{\lambda}_{CPCSW,W\_CPC}$ were both close to 0.8 (e.g., $\hat{\lambda}^{2MLR}_{CPCSW,W\_CPC} = 0.358$, 95% = 0.944, Sig = 0.796; $\hat{\lambda}^{2MaxB}_{CPCSW,W\_CPC}$, 95% = 0.951, Sig = 0.819). In the true model, these two factor loadings were considered as non-zero and smaller effects without statistical significance at small sample size. With the small sample size setting in the simulation, this kind of smaller effects were set to have less empirical rate of significant estimates over total replications than the nominal level of 0.8 (Eng, 2003). Results of the 2MLR and 2MaxB were consistent with the population model, in which only the empirical power for the factor loading of individual-level athletic competence (CPCAC) far more than 0.8 but not those for social acceptance (CPCSA) and self-worth (CPCSW).

Variance Explained[5] and Scale Reliability of Indicators

Taking [CN(CS), ICC] = [40(3), 0.3] as an example shown in **Table 5**, 1MLR tended to generate inflated R^2 measure, especially for the indicators with smaller within-level factor loadings but larger between-level factor loadings, so did the 2Miss model (e.g., $\hat{R}^{2,\ 1MLR}_{CPCSA} = 0.467$ and, $\hat{R}^{2,\ 2Miss}_{CPCSA,\ Within-level} = 0.278$ vs. $\hat{R}^{2,\ 2MLR}_{CPCSA,\ Within-level} = 0.175$ and $\hat{R}^{2,\ 2MaxB}_{CPCSA,\ Within-level} = 0.176$). As for the between-level, 2MaxW provided consistent $\hat{R}^2$ as 2MLR but 2Miss generated biased estimate ($\hat{R}^{2,\ 2MaxW}_{CPCSA,\ Between-level} =$

[3] The 95% confidence interval coverage rate (95%, defined as the empirical proportion for which the 95% confidence interval of estimate contained the true population parameter value).

[4] Empirical power (Sig., defined as the empirical significance pattern of the estimates; that is empirical power = average rate of significant estimates over total replications).

[5] In two-level models, the variables are partitioned into level-l and level-2 components. So the R^2 computed in these approaches should be interpreted as the proportion of variance in each within-group component that is accounted for by the lower-level model, and the proportion of variance in each between-group component that is accounted for by the higher-level model; while in 1-level MLR, R^2 is proportion of variance in each indicators that is accounted for by an overall model where the variance composition is confounded by components from both levels.

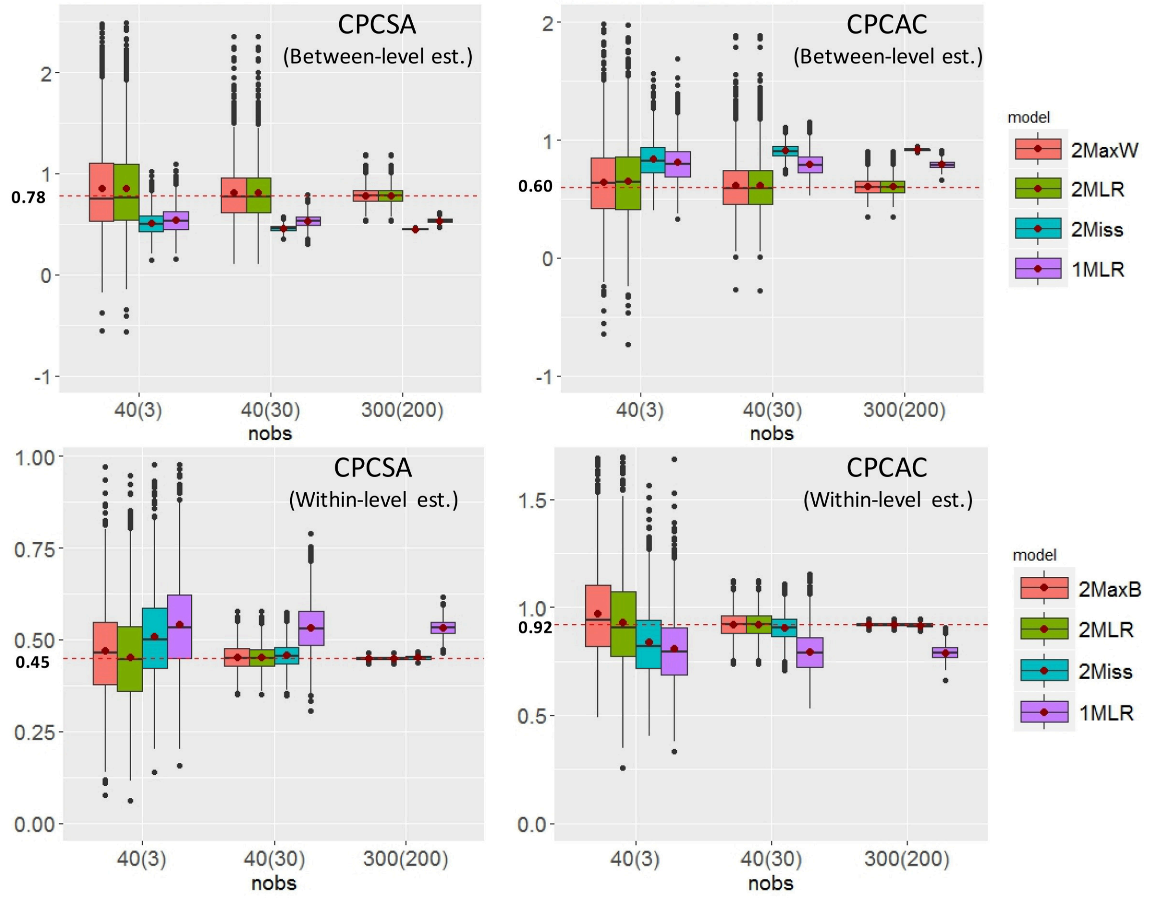


FIGURE 4 | The Boxplots of selected factor loading estimates vs. sample size conditions. The red dots in the boxes indicate the means of factor loading estimates. The red dashed lines indicates the parameter settings in respective levels.

0.746 and $\hat{R}^{2,\ 2MLR}_{CPCSA,\ Between-level} = 0.740$ vs. $\hat{R}^{2,\ 2Miss}_{CPCSA,\ Between-level} = 0.623$).

As for the scale reliability, the 2MaxB and 2MaxW yielded consistent reliability measures as 2MLR in respective levels, but 1-level MLR and 2-level Miss tended to underestimate the score consistency of indicators (e.g., $\hat{\rho}^{2MaxB}_{Within-level} = 0.830 \cong \hat{\rho}^{2MLR}_{Within-level} = 0.825$; $\hat{\rho}^{2MaxW}_{Between-level} = 0.926 \cong \hat{\rho}^{2MLR}_{Between-level} = 0.930$ vs. $\hat{\rho}^{1MLR} = 0.747$, $\hat{\rho}^{2Miss}_{Within-level} = 0.798$ and $\hat{\rho}^{2Miss}_{Between-level} = 0.915$).

In summary, given the conflated estimates of fixed and random effects, the 1MLR models would provide overestimated variance explained measure and underestimated reliability measure for the indicators. In contrast, the 2MaxB and 2MaxW model generated consistent R^2 and ρ for respective within-level and between-level indicators consistent with those of the 2MLR model across simulation settings.

DISCUSSION AND CONCLUSION

As researchers call for the need to adequately take into account of the multilevel structure of social and behavioral data (Skinner et al., 1997; Lee and Forthofer, 2006), the use of multilevel data modeling techniques will be inevitable. However, multilevel models are not an infallible statistical strategy unless the hypothesized model conforms to the real data structure. In this study, we demonstrated that maximum models are robust analytic methods as to the inequality of higher- and lower-level factor loadings or to detect possibly non-significant trivial items, especially when researchers have limited information about the significance pattern of factor loadings and level-varying measurement structures. The current study focuses on multilevel CFA, which is a generic form of structural equation models; therefore, the study result can be generalized to more complex models.

Specifically, we examined the performance of five proposed SEM techniques on analyzing complex survey data with unequal factor loadings under equal between- and within-level structures. Across different combinations of cluster numbers, cluster sizes and ICC values, all models yield acceptable model-fit information. AIC and adjusted BIC could be utilized to differentiate 1MLR from 2-level models but could not select the best 2-level model. Among 2-level models, 2MLR, 2MaxB and 2MaxW could consistently generate the effective and efficient parameter estimates. On the contrary, the design-based model would not be an appropriate approach on analyzing complex

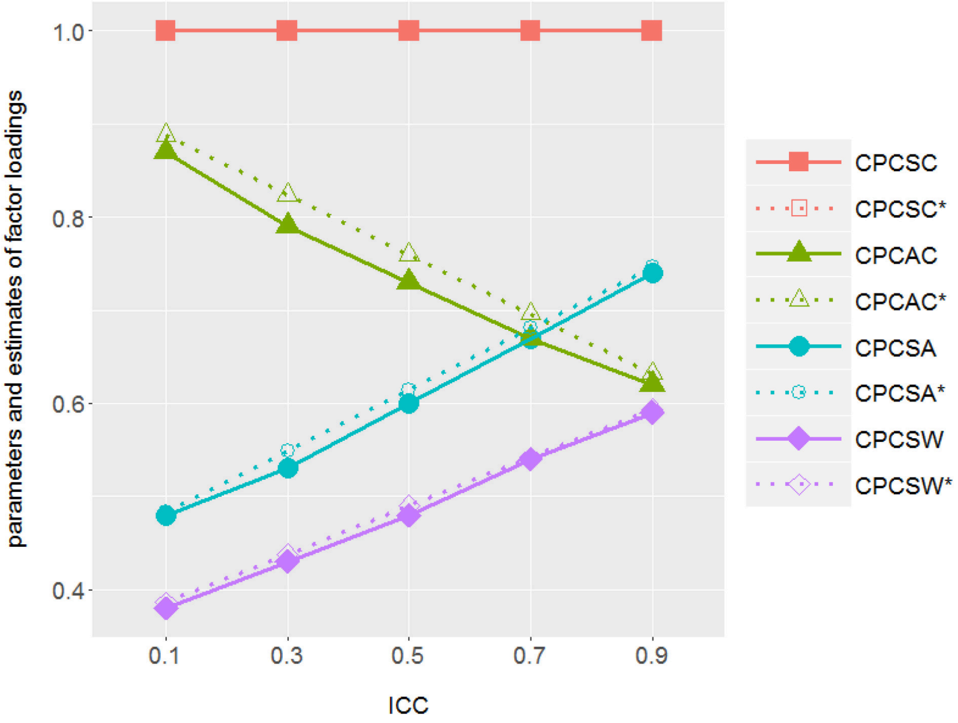


FIGURE 5 | ICCs vs. parameter estimates from the simulations and those from the mathematical derivations of the design-based approach: As ICC increases, design-based approach tends to generate factor loading estimates which are closer to its between-level counterpart and deviate from its within-level values in the true model. There is one factor in both within and between levels with factor variance $\boldsymbol{\Psi}_{Within-level} = (\mathbf{1} - \mathbf{ICC})\cdot\boldsymbol{\Psi}_{total}$ and $\boldsymbol{\Psi}_{Between-level} = \mathbf{ICC}\cdot\boldsymbol{\Psi}_{total}$ with $\boldsymbol{\Psi}_{total} = \boldsymbol{\Psi}_{Between-level} + \boldsymbol{\Psi}_{Within-level} = 1$. Solid line illustrates the factor loading estimates of design-based approach (1MLR) from simulations; dotted line illustrates the theoretical parameter values of design-based approach. CPCSC is the marker variable. The true value of CPCAC $\lambda^{True\ model}_{CPCAC,\ W\_CPC} = 0.920$, $\lambda^{True\ model}_{CPCAC,\ B\_CPC} = 0.600$; CPCSA $\lambda^{True\ model}_{CPCSA,\ W\_CPC} = 0.450$ and $\lambda^{True\ model}_{CPCSA,\ B\_CPC} = 0.780$; CPCSA $\lambda^{True\ model}_{CPCSW,\ W\_CPC} = 0.360$ and $\lambda^{True\ model}_{CPCSW,\ B\_CPC} = 0.620$.

TABLE 5 | Values of ICC and R^2 on indicators in the synthetic dataset of harter's competence measures using five SEM modeling techniques for [CN(CS), ICC] = [40(3), 0.3].

			CPCSC	CPCSA	CPCAC	CPCSW	Scale reliability$_\rho$
ICC			0.617	0.612	0.352	0.431	---
R^2	2MLR	Within-level	0.506	0.175	0.473	0.124	0.825
		Between-level	0.802	0.740	0.651	0.662	0.930
	1MLR		0.697	0.467	0.482	0.370	0.747
	2MaxB	Within-level	0.503	0.176	0.473	0.124	0.830
	2MaxW	Between-level	0.799	0.746	0.653	0.660	0.926
	2Miss	Within-level	0.468	0.278	0.271	0.195	0.798
		Between-level	0.843	0.623	0.745	0.553	0.915

CPCSC, Harter perceived scholastic competence; CPCSA, Harter perceived social acceptance; CPCAC, Harter perceived athletic competence; CPCSW, Harter perceived global self-worth.

survey data due to its conflated fixed and random effect estimates, inflated standard error estimates, and inconsistent statistical inferences, along with the overestimated variance explained and underestimated reliability measures of the indicators. Below we elaborated on the consequences of using design-based models and miss-specified 2-level models as well as the advantages of our recommended methods in analyzing complex survey measurement data.

Disadvantages of the Design-Based Approach and Mis-Specified Multilevel Models

Using both mathematical derivation and empirical data simulation, we demonstrated that the 1MLR as well as 2Miss yields similar but conflated fixed effect; on the other hand, 2Miss could specify level-specific random components while 1MLR would yield overall random effect estimates. When 1MLR model is used, it truly estimates the combination of variations from different levels in a single-level modeling simultaneously. The parameter estimates got mixed with components from both levels except for ICC = 0 and 1 (as shown in **Figure 5**). In that case, the consequences were spurious fixed effect estimates with more likely statistical significance and bigger R^2. Moreover, with the overall estimate of residual variance, the design-based approach tended to generate smaller scale reliability estimates.

On the other hand, if the model-based approach is miss-specified, researchers will yield parameter estimates which deviate from the population values in respective levels. In this study, we construct the miss-specified model-based model by constraining the between- and within-level factor loadings to be equal, and the consequence of the analytic results is similar to that of the design-based approach because the design-based approach assumes the between and within level model have not only exactly the same structure (Muthén and Satorra, 1995; Wu and Kwok, 2012), but also the same magnitude of factor loadings.

In regression-like analyses, the design-based approach is reliable to generate consistent statistical inference of parameter estimates by adjusting its standard error considering data dependency (Hardin and Hilbe, 2007); however, in CFA or SEM-based analysis, we demonstrate that the design-based approach on complex survey data cannot guarantee consistent statistical inferences of the result to a specific level with conflated parameter estimates. Design-based approaches are beneficial to take the data dependency into consideration by adjusting the estimate of standard error when the between and within levels have equal structures. However, only when the equality in structures and in population values holds for both levels, the analytic result can be unbiased to specific-level inferences. In most of MCFA or MSEM analyses, the parameter estimates obtained from the design-based approach is a function of between- and within-level population values and the analytic result cannot infer to any level. In the case of children's perceived Harter competence, the four factor loadings of different competence aspects were all statistically significant at the classroom level while only the factor loading of athletic competence was significant at the individual level in early childhood, based on a correctly specified and analyzed result. Nevertheless, as shown in the 1MLR and 2Miss, all four factor loadings were statistically significant which could mislead researchers to conclude that all four competence aspects were important for the *individual* development of the overall perceived competence and to invest their efforts to items (aspects) that are trivial or of little importance for early elementary student's individual competence development.

Under the MCFA framework, we provided evidence to illustrate that design-based approaches yield conflated parameter estimates with multilevel measurement data even under equal level structures as long as the population values at each level are different. In reality, we can hardly know the true model and thus should be more cautious about making inferences with estimates from design-based approaches to represent the lower-level model characteristics.

Advantages of Maximum Models

To have consistent and unbiased statistical inferences, methodologists debated over the adequacy of model-based approaches and design-based approaches on analyzing multilevel dataset from complex survey sampling (Snijders and Bosker, 2011). Adding new findings to the literature, first, we demonstrated that the design-based approach is not a robust analytic model for multilevel data under equal level structures with unequal factor loadings. Second, the model-based approach can produce unbiased fixed and random effect estimates as well as their corresponding statistical inferences if and only if the model is correctly specified. Third, most importantly, we suggested that 2MaxB and 2MaxW models are robust and feasible techniques for separating variance components from different levels and for investigating possible higher-level and lower-level structures. Fourth, when the number of clusters in the higher-level sampling units is sufficient (e.g., no less than 40 as shown in simulation), the 2MLR and 2MaxW models can yield consistent and effective estimates of the fixed and random effects. By estimating a saturated between- or within-level model, maximum models enable researchers to focus on examining the lower- or higher-level findings and to obtain consistent statistical inference for findings that researchers are interested in. In the current empirical data simulation, compared to those in the design-based model, variables with smaller factor loadings and smaller R^2 in the within level of the maximum model (e.g., social competence in 2MaxB model) may suggest stronger factor loadings in the between level based on the Equation (7). Researchers in the applied area are encouraged to compare results from maximum models with those from design-based models to investigate possible higher level variation and avoid investing unnecessary efforts on unimportant aspects (i.e., trivial items with smaller amount of factor loadings and variance explained).

Recommendations for Practice and Limitation

According to the simulation results, information criteria performed better than model-fit test and fit indices in selecting the optimal analytical models on multilevel measurement data. Researchers can refer to information criteria statistics to determine if their hypothesized models fit the multilevel measurement data adequately. They can start by fitting a 2Miss and a 1MLR. If the information criteria suggested better fit for the 2Miss model (e.g., Δ *AIC* or Δ *adj.BIC* $\geq$ 20), they should go a step further to perform 2MLR when they have theoretical or empirical evidence, or they could specify 2MaxB or 2MaxW depending on their primary interest in the specific level to ensure consistent and effective estimates of the fixed and random effects. Especially 2MaxB is recommended when the number of between-level sampling units is small (e.g., CN < 40) under the setting of 4 or fewer manifest variables. As a caveat, though AIC and adj. BIC reflected better fit for 2-level models than design-based models across all simulation conditions, they were shown to perform poorly in many contexts (e.g., Preacher and Merkle, 2012). More research can be done to investigate the effectiveness of AIC and BIC in model selection across different parameter settings.

Moreover, in this study, we discuss a multilevel measurement model with a uni-factor structure in both levels; however, if the level structure is misspecified, part of the misspecification would still pass on to the other level and influence the modeling result. Thus, it is possible that the residual part may not truly reflect the misspecification in 2MaxW or 2MaxB. Similar concerns have been raised for developing the method of MUML (Muthén, 1994) and for separately evaluating the within and between level structures (Yuan and Bentler, 2007). Since it is very unlikely to have a correct model specification in practice, results obtained for 2MaxB and 2MaxW may be too optimistic to generalize for empirical dataset. The performance of 2MaxB and 2MaxW models applied in substantial research warrants for future investigation. In addition, the model specification may become more complicated when there is more than one factor or when the observed variables are not normally distributed. Future study can be conducted to investigate the performance of 2MaxB and 2MaxW in more complex settings.

AUTHOR CONTRIBUTIONS

JW designed the study, conducted the simulation study and took a leading role in writing the manuscript. JL, MN, and YH conducted a part of the simulation and prepared some tables and figures.

FUNDING

This research was partially supported by Grant 104-2628-H-009-001-SS3 and 105-2511-S-009-009-MY3 from the Ministry of Science and Technology, Taiwan.

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Contr.* 19, 716–723. doi: 10.1109/TAC.1974.1100705

Bagozzi, R. P., and Yi, Y. (1988). On the evaluation of structural equation models. *J. Acad. Mark. Sci.* 16, 74–94. doi: 10.1007/BF02723327

Bollen, K. A., Harden, J. J., Ray, S., and Zavisca, J. (2014). BIC and alternative Bayesian information criteria in the selection of structural equation models. *Struct. Eq. Model.* 21, 1–19. doi: 10.1080/10705511.2014.856691

Brook, D. W., Rubenstone, E., Zhang, C., Morojele, N. K., and Brook, J. S. (2011). Environmental stressors, low well-being, smoking, and alcohol use among South African adolescents. *Soc. Sci. Med.* 72, 1447–1453. doi: 10.1016/j.socscimed.2011.02.041

Brown, T. A. (2006). *Confirmatory Factor Analysis for Applied Research, 1st Edn.* New York, NY: The Guilford Press.

du Toit, S. H., and du Toit, M. (2008). "Multilevel structural equation modeling," in *Handbook of Multilevel Analysis*, eds J. de Leeuw and E. Meijer (New York, NY: Springer), 435–478.

Dyer, N. G., Hanges, P. J., and Hall, R. J. (2005). Applying multilevel confirmatory factor analysis techniques to the study of leadership. *Leadersh. Q.* 16, 149–167. doi: 10.1016/j.leaqua.2004.09.009

Eng, J. (2003). Sample size estimation: how many individuals should be studied? *Radiology* 227, 309–313. doi: 10.1148/radiol.2272012051

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

Geldhof, G. J., Preacher, K. J., and Zyphur, M. J. (2013). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychol. Methods*. 19, 72–91. doi: 10.1037/a0032138.

Hardin, J. W., and Hilbe, J. M. (2007). *Generalized Linear Models and Extensions, 2nd Edn.* College Station, TX: Stata Press.

Harter, S. (1982). The perceived competence scale for children. *Child Dev.* 53, 87–97. doi: 10.2307/1129640

Heck, R. H., and Thomas, S. L. (2008). *An Introduction to Multilevel Modeling Techniques, 2nd Edn.* New York, NY: Routledge.

Hox, J. J. (2002). *Multilevel Analysis Techniques and Applications*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hox, J. J. (2010). *Multilevel Analysis: Techniques and Applications, 2nd Edn.* New York, NY: Routledge Academic.

Hsu, H.-Y., Kwok, O., Lin, J. H., and Acosta, S. (2015). Detecting misspecified multilevel structural equation models with common fit indices: a monte carlo study. *Multiv. Behav. Res.* 50, 197–215. doi: 10.1080/00273171.2014.977429

Hu, L., and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struc. Eqn Model.* 6, 1–55. doi: 10.1080/10705519909540118

Huber, P. J. (1967). "The behavior of maximum likelihood estimates under nonstandard conditions," in *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, eds L. M. Le Cam and J. Neyman (Berkeley, CA: Statistical Laboratory of the University of California; University of California Press), 221–233.

Hughes, J., and Kwok, O. (2007). Influence of student-teacher and parent-teacher relationships on lower achieving readers' engagement and achievement in the primary grades. *J. Educ. Psychol.* 99, 39–51. doi: 10.1037/0022-0663.99.1.39

Klangphahol, K., Traiwichitkhum, D., and Kanchanawasi, S. (2010). Applying multilevel confirmatory factor analysis techniques to perceived homework quality. *Res. Higher Edu.* 6, 1–10.

Kline, R. B. (2010). *Principles and Practice of Structural Equation Modeling, 3rd Edn.* New York, NY: The Guilford Press.

Lee, E. S., and Forthofer, R. N. (2006). *Analyzing Complex Survey Data*. Newbury Park, CA: Sage.

MacCallum, R. C., Widaman, K. F., Zhang, S., and Hong, S. (1999). Sample size in factor analysis. *Psychol. Methods* 4, 84–99. doi: 10.1037/1082-989X.4.1.84

Marsh, H. W., and Hau, K. T. (2007). Applications of latent-variable models in educational psychology: the need for methodological-substantive synergies. *Contemp. Educ. Psychol.* 32, 151–170. doi: 10.1016/j.cedpsych.2006.10.008

Martin, M. J., McCarthy, B., Conger, R. D., Gibbons, F. X., Simons, R. L., Cutrona, C. E., et al. (2011). The enduring significance of racism: discrimination and delinquency among black American youth. *J. Res. Adolesc.* 21, 662–676. doi: 10.1111/j.1532-7795.2010.00699.x

Meade, A. W., and Bauer, D. J. (2007). Power and precision in confirmatory factor analytic tests of measurement invariance. *Struct. Eq. Model.* 14, 611–635. doi: 10.1080/10705510701575461

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *J. Educ. Measur.* 28, 338–354. doi: 10.1111/j.1745-3984.1991.tb00363.x

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociol. Methods Res.* 22, 376–398. doi: 10.1177/0049124194022003006

Muthén, B. O., and Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociol. Methodol.* 25, 267–316. doi: 10.2307/271070

Muthén, L. K., and Muthén, B. O. (2012). *Mplus User's Guide ,7th Edn.* Los Angeles, CA: Muthén and Muthén.

Nylund, K. L., Asparouhov, T., and Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Struct. Eq. Model.* 14, 535–569. doi: 10.1080/10705510701575396

Preacher, K. J., and Merkle, E. C. (2012). The problem of model selection uncertainty in structural equation modeling. *Psychol. Methods* 17, 1–14. doi: 10.1037/a0026804

Rabe-Hesketh, S., and Skrondal, A. (2008). *Multilevel and Longitudinal Modeling Using Stata*. College Station, TX: Stata Press.

Raykov, T. (2004). Behavioral scale reliability and measurement invariance evaluation using latent variable modeling. *Behav. Ther.* 35, 299–331. doi: 10.1016/S0005-7894(04)80041-8

Rebollo, I., de Moor, M. H. M., Dolan, C. V., and Boomsma, D. I. (2006). Phenotypic factor analysis of family data: correction of the bias due to dependency. *Twin Res. Hum. Genet.* 9, 367–376. doi: 10.1375/twin.9.3.367

Roberts, G., Mohammed, S. S., and Vaughn, S. (2010). Reading achievement across three language groups: growth estimates for overall reading and reading subskills obtained with the early childhood longitudinal survey. *J. Educ. Psychol.* 102, 668–686. doi: 10.1037/a0018983

Róbert, U. (2010). Early smoking experience in adolescents. *Addict. Behav.* 35, 612–615. doi: 10.1016/j.addbeh.2009.12.018

Rosenthal, J. A., and Villegas, S. (2010). Living situation and placement change and children's behavior. *Child. Youth Serv. Rev.* 32, 1648–1655. doi: 10.1016/j.childyouth.2010.07.003

Ryu, E., and West, S. G. (2009). Level-specific evaluation of model fit in multilevel structural equation modeling. *Struc. Eq. Model.* 16, 583–601. doi: 10.1080/10705510903203466

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.* 6, 461–464. doi: 10.1214/aos/1176344136

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika* 52, 333–343. doi: 10.1007/BF02294360

Skinner, C., Holt, D., and Wrigley, N. (1997). *The Analysis of Complex Survey Data*. Hoboken, NJ: John Wiley and Sons Inc.

Skrondal, A., and Rabe-Hesketh, S. (2007). Latent variable modelling: a survey. *Scand. J. Stat.* 34, 712–745. doi: 10.1111/j.1467-9469.2007.00573.x

Snijders, T. A. B., and Bosker, R. (2011). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling, 2nd Edn*. London: SAGE Publications Ltd.

Stapleton, L. M. (2006). "Using multilevel structural equation modeling techniques with complex sample data," in *Structural Equation Modeling: A Second Course*, eds G. R. Hancock and R. O. Mueller (Greenwich, CT: Information Age Publishing), 345–383.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 48, 817–838. doi: 10.2307/1912934

Wilhelm, P., and Schoebi, D. (2007). Assessing mood in daily life: structural validity, sensitivity to change, and reliability of a short-scale to measure three basic dimensions of mood. *Eur. J. Psychol. Assess.* 23, 258–267. doi: 10.1027/1015-5759.23.4.258

Wu, J.-Y. (2015). University students' motivated attention and use of regulation strategies on social media. *Comput. Educ.* 89, 75–90. doi: 10.1016/j.compedu.2015.08.016

Wu, J.-Y. (2017). The indirect relationship of media multitasking self-efficacy on learning performance within the personal learning environment: implications from the mechanism of perceived attention problems and self-regulation strategies. *Comput. Educ.* 106, 56–72. doi: 10.1016/j.compedu.2016.10.010

Wu, J.-Y., Hughes, J. N., and Kwok, O. (2010). Teacher-student relationship quality type in elementary grades: effects on trajectories for achievement and engagement. *J. Sch. Psychol.* 48, 357–387. doi: 10.1016/j.jsp.2010.06.004

Wu, J.-Y., and Kwok, O. (2012). Using structural equation modeling to analyze complex survey data: a comparison between design-based single-level and model-based multi-level approaches. *Struct. Eq. Model.* 19, 16–35. doi: 10.1080/10705511.2012.634703

Wu, J.-Y., Kwok, O., and Willson, V. L. (2014). Using design-based latent growth curve modeling with cluster-level predictor to address dependency. *J. Exp. Educ.* 82, 431–454. doi: 10.1080/00220973.2013.876226

Yuan, K.-H., and Bentler, P. M. (2007). Multilevel covariance structure analysis by fitting multiple single-level models. *Sociol. Methodol.* 37, 53–82. doi: 10.1111/j.1467-9531.2007.00182.x

Yuan, K.-H., Hayashi, K., and Bentler, P. M. (2007). Normal theory likelihood ratio statistic for mean and covariance structure analysis under alternative hypotheses. *J. Multivar. Anal.* 98, 1262–1282. doi: 10.1016/j.jmva.2006.08.005

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

APPENDIX

Mplus Syntax for 2MaxB Model

```
TITLE:     This is an example of a Maximum model
DATA:      FILE = Harter3_change_21.dat;
VARIABLE:  NAME = cpc31-cpc34 Cluster;
           USEVARIABLES = cpc31-cpc34;
           CLUSTER = Cluster;
ANALYSIS:  TYPE = TWOLEVEL;
MODEL:
         %Within%                                           ! Set up Within-level Model
         cpc3w BY cpc31@1 cpc32 cpc33                       ! Specify lower-level CFA model
         cpc34;                                             ! Item residual variance
         cpc31 cpc32 cpc33 cpc34;                           estimates
         %Between%                                          ! Set up Between-level Model
         cpc31 WITH cpc32 cpc33 cpc34;                      ! Estimate full rank
         cpc32 WITH cpc33 cpc34;                            ! variance-covariance matrix in
         cpc33 WITH cpc34;                                  ! higher-level structure
         [cpc31 cpc32 cpc33 cpc34];                         ! Item intercept estimates

OUTPUT:  SAMP RES STAND MOD;
```

frontiers in Applied Mathematics and Statistics

ORIGINAL RESEARCH
published: 25 October 2017
doi: 10.3389/fams.2017.00019

Specifying Turning Point in Piecewise Growth Curve Models: Challenges and Solutions

Ling Ning[1]* *and Wen Luo*[2]

[1] *Center for Student Affairs Assessment, University of California, Davis, Davis, CA, United States,* [2] *Texas A&M University, College Station, TX, United States*

Piecewise growth curve model (PGCM) is often used when the underlying growth process is not linear and is hypothesized to consist of phasic developments connected by turning points (or knots or change points). When fitting a PGCM, the conventional practice is to specify turning points a priori. However, the true turning points are often unknown and misspecifications of turning points may occur. The study examined the consequences of turning point misspecifications on growth parameter estimates and evaluated the performance of commonly used fit indices in detecting model misspecification due to mis-specified locations of turning points. In addition, this study introduced and evaluated a newly developed PGCM which allows unknown turning points to be freely estimated. The study found that there are severe consequences of turning point misspecification. Commonly used model fit indices have low power in detecting turning point misspecification. On the other hand, the newly developed PGCM with freely estimated unknown turning point performs well in general.

Keywords: latent growth curve model, piecewise, turning point, model fit indices, MI

OPEN ACCESS

Edited by:
Suzanne Jak,
University of Amsterdam, Netherlands

Reviewed by:
Antonio Calcagnì,
University of Trento, Italy
Kees-Jan Kan,
University of Amsterdam, Netherlands

****Correspondence:***
Ling Ning
lning@ucdavis.edu

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Applied Mathematics and Statistics

Received: *19 May 2017*
Accepted: *02 October 2017*
Published: *25 October 2017*

Citation:
Ning L and Luo W (2017) Specifying Turning Point in Piecewise Growth Curve Models: Challenges and Solutions.
Front. Appl. Math. Stat. 3:19.
doi: 10.3389/fams.2017.00019

INTRODUCTION

Longitudinal studies have been widely applied in many research areas to examine individual differences in growth over time. One commonly used method to study individual change over time is the latent growth modeling in the Structural Equation Modeling (SEM) framework [1, 2]. Up to date, the majority of applications of the latent growth models in longitudinal data analyses have been limited to the assumption that the change follows a simple linear trend. However, when longitudinal data are collected over an adequately long period of time, the features of individual change do not always follow a linear trend.

A more flexible approach to model the nonlinear form of growth is the piecewise growth curve model (PGCM). This approach breaks up the curvilinear growth trend into separate linear segments or pieces of different slopes, which are tied together by turning points (or knots or change points). The flexibility of PGCM allows the formulation of different functional forms for the different phases of growth such that each phase does not have to conform to the same function [3–6]. The approach is particularly appealing when researchers are interested in comparing growth rates for two or more periods, such as the effect of schooling on children's scholastic attainments before and after secondary school [7, 8].

The major difficulty in applying PGCMs concerns the specification of the turning point. Researchers tend to rely on theories or designs (e.g., the start point of an intervention) to choose the location of the turning point (see e.g., [9, 10]). Yet, such considerations may not always be reasonable. For example, the turning point may occur after the intervention due to delay in response to intervention. The misspecification of a turning point may render a suboptimal functional representation of the observed data patterns, leading to incorrect inferences of growth traits.

Alternative approaches were developed to search for the optimal location of the turning point based on data [6, 11, 12]. For example, Kwok et al. [6] proposed using modification index to detect the turning point in the linear latent growth modeling framework. Harring et al. [3] extended PGCM to treat the turning point as an unknown parameter to be estimated in the SEM framework. Compared to the conventional PGCM with turning points specified a priori, such an extension is appealing because researchers do not have to have a priori knowledge of the turning points. Moreover, allowing for free estimation of turning points and time specific factor loadings can lead to a more optimal functional form of each growth phase, giving a more adequate description of the growth pattern in the data [6, 13]. The appealing advantages of the newly proposed PGCM with unknown turning points have attracted an increasing amount of interest in empirical studies (see e.g., [5, 14, 15]).

Comparing and contrasting the conventional and the new PGCM, this study aims to investigate the three research questions. First, under what conditions and to what extent does the misspecification of turning point in conventional PGCMs have a substantial impact on the growth trait estimation? Second, in conventional PGCMs, can the commonly used fit indices correctly identify model misspecification due to the mislocation of turning points? Lastly, can the new procedure of PGCM with an unknown turning point accurately estimate the turning point and growth parameters?

The remaining of the paper is organized in the following sections. We first reviewed the model specification for the new PGCM with an unknown turning point, followed by a brief description of commonly used fit indexes under the SEM framework. Then we introduced the methods for data generation, analysis procedure, and presented the findings from the simulation study. Finally, we discussed the findings in relation to previous studies, implications, and limitations.

PGCM WITH ONE UNKNOWN TURNING POINT

Suppose that the sample data consist of *j* equal spaced repeated measures of **Y** for individual *i*. A two-piece growth model with one unknown turning point can be specified in the form of two-level models. The Level 1 (repeated measures) model is specified as

$$y_{ij} = \begin{cases} l_1(t) : a_{1i} + b_{1i}(t_{ij}) + \varepsilon_{ij} & t_{ij} \leqslant \gamma \\ l_2(t) : a_{2i} + b_{2i}(t_{ij}) + \varepsilon_{ij} & t_{ij} > \gamma \end{cases} \tag{1}$$

where y_{ij} is the response at the *j*th measurement for the *i*th individual. a_{1i} and b_{1i} are the intercept and the slope growth factors before the occurrence of the turning point, and a_{2i} and b_{2i} denote the corresponding growth factors after the turning point. γ is the location of the turning point marking the shift from one growth phase to the other. ε_{ij} is the level-1 residual for individual *i* at measurement *j* $[\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)]$. It is assumed that the location of the turning point is fixed to be the same for all individuals. Hence the model is appropriate when homogeneous turning points are assumed. For example, studies have found that almost all the average children have been able to establish their numerical and arithmetic foundation in 3rd grade, which could be assumed to be a common turning point in the development of child numerical cognition (see e.g., [16]).

The trajectory is assumed to be continuous and has no gap between the two pieces, such that the two pieces for $l_1(t)$ and $l_2(t)$ are connected at the turning point. That is, when $t_{ij} = \gamma$, $a_{1i} + b_{1i}(\gamma) = a_{2i} + b_{2i}(\gamma)$, which gives $a_{2i} = a_{1i} + \gamma(b_{1i} - b_{2i})$. Thus Model (Equation 1) that has five parameters is reduced to a four-parameter model

$$y_{ij} = \begin{cases} l_1(t) : a_{1i} + b_{1i}(t_{ij}) + \varepsilon_{ij} & t_{ij} \leqslant \gamma \\ l_2(t) : a_{1i} + b_{1i}\gamma + b_{2i}(t_{ij} - \gamma) + \varepsilon_{ij} & t_{ij} > \gamma \end{cases} \tag{2}$$

The Level-2 (between-subject) model is specified as

$$\begin{cases} a_{1i} = \mu_{a1} + \zeta_{a1i} \\ b_{1i} = \mu_{b1} + \zeta_{b1i} \\ b_{2i} = \mu_{b2} + \zeta_{b2i} \end{cases} \tag{3}$$

with

$$\begin{bmatrix} \zeta_{a1i} \\ \zeta_{b1i} \\ \zeta_{b2i} \end{bmatrix} \sim \text{MVN}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{\pi 00} & \tau_{\pi 01} & \tau_{\pi 02} \\ \tau_{\pi 10} & \tau_{\pi 11} & \tau_{\pi 12} \\ \tau_{\pi 20} & \tau_{\pi 21} & \tau_{\pi 22} \end{bmatrix} \right), \tag{4}$$

where μ_{a1}, μ_{b1}, and μ_{b2} are growth factor means and ζ_{a1}, ζ_{b1}, and ζ_{b2} are random disturbances in their respective growth factors. The Level 1 residuals and the Level 2 disturbances are also assumed to be uncorrelated with each other and with the latent growth factors.

The parameterization of Model (Equation 2) cannot be specified and estimated directly in conventional Structural Equation Modeling (SEM) programs. Harring et al. [3] suggested a re-parameterization of Model (Equation 2) to make the estimation in SEM programs possible. They proposed to combine the two linear trajectories in Model (Equation 2) into one equation $y_{ij} = \lambda_{1i} + \lambda_{2i}t_{ij} + \lambda_{3i}\sqrt{(t_{ij} - \gamma)^2} + \varepsilon_{ij}$, where $\lambda_{1i} = (a_{1i} + a_{2i})/2$, $\lambda_{2i} = (b_{1i} + b_{2i})/2$, and $\lambda_{3i} = (b_{2i} - b_{1i})/2$. Readers are referred to Harring et al. [3] and Kohli and Harring [5] for details of the model re-parameterization.

SEM-BASED FIT INDICES

The commonly used fit indices available in standard SEM software for applied researchers to determine the adequacy of their SEM models includes but not limited to root-mean-square error of approximation (RMSEA; [17]), standardized root-mean-square residual (SRMR; [18]), the Comparative Fit Index (CFI; [19]), and the Tucker-Lewis Index (TLI; [20]). Following the recommendation of Hu and Bentler [21, 22], the cutoff criteria for the commonly used fit indices (e.g., RMSEA $\leq$ 0.06; CFI $\geq$ 0.95; TLI $\geq$ 0.95; SRMR $\leq$ 0.08) have been generally used to assess model fit/misfit in SEM analysis. However, there has been controversy regarding the advocacy for the proposed fixed cutoff criteria. Applied researchers were warned against a complete reliance on fixed cutoff criteria in assessing model fit (see e.g., [23–25]). Simulation studies have been done in the context of confirmatory factor analysis (CFA) models, evaluating the performance of the fit indices in identifying misspecification in covariance structures (see e.g., [21, 23, 24]).

More relevant to the present research interest were the studies that addressed the sensitivity of fit indices in identifying misspecifications of growth shape. Wu et al. [26] derived theoretically that the SEM-based fit indices such as the Chi-Square, RMSEA, CFI and TLI were able to "directly detect" mis-specified functional form for the mean growth trajectory. Wu and West [27] evaluated the theoretical derivation using a simulation study to further understand the performance of the above mentioned fit indices in detecting model misspecification in covariance structures and marginal mean structure. In their study, the mean growth trajectory in the population model was quadratic GCM, but was mis-specified as linear GCM. Their findings with regards to the capabilities of fit indices in detecting mis-specified mean functional forms showed that RMSEA, CFI, and TLI were more sensitive to misspecification in marginal mean structure than Chi-square test statistic or SRMR, while the latter two were affected by sample size. Leite and Stapleton [28] found that comparatively speaking, the Chi-square test statistic performed the best, followed by RMSEA, relative to CFI, TLI, and SRMR in detecting model misfit in GCM, accounting for sample size, misspecification severity, number of time points, and population growth shapes, when the population data generated using quadratic, plateau, and piecewise GCMs were fitted using a (mis-specified) linear model. It is noteworthy that the baseline model used to calculate CFI and TLI for growth curve model is not appropriate in standard SEM software packages including the Mplus software. The appropriate baseline model is an intercept-only model in which only the intercept mean and residual variances are freely estimated [27, 29].

Another important piece of information that applied researchers tend to rely on for model fit improvement is modification index (MI) or Lagrange multiplier. What MI captures is an estimate of the expected change in the specified model's overall chi-square (χ^2) value if a previously constrained parameter were allowed to be freely estimated. A large MI value suggests an appreciable improvement in model fit if the model were modified to freely estimate that particular parameter, given that the post hoc modification is theoretically justifiable. While MI provides the significance of the misspecification, EPC (expected parameter change) is an estimate of the impact of the misspecification on parameter estimates. EPC has been suggested to be used in conjunction with MI to detect model misspecification (see e.g., [30]). Several variations of EPC have been proposed: the unstandardized expected parameter change (EPC; [31]), which provides the estimated value that a given fixed parameter would have if it were freely estimated in the model; the partially standardized EPC [32], and the fully standardized EPC (SEPC; [33]), referred to as "Std YX E.P.C." in the Mplus package (2007–2016). Interested readers are referred to Whittaker [34] for the differences between the variations of EPC.

Saris et al. [31] argued against the reliance on χ^2 test statistics and fit indices for model evaluation because they are not only affected by the degree of misspecification but also by the incidental characteristics of the model. Alternatively, they proposed to use MI along with EPCs. However, the decision on the presence of model misspecification can only be made when a large, significant MI is associated with a large EPC. Saris et al. [35] further suggested taking into account of the information on the power of the MI test when using the SEPC in combination with MI to make decision regarding model misspecification errors. They also suggested that a SEPC of 0.2 or larger is a large value, indicative of possible misspecification error. To evaluate whether a SEPC of 0.2 or larger can be implemented as a cutoff criterion of the SEPC in applied research, Whittaker [34] conducted a simulation study to examine the performance of the MI and SEPC in detecting misspecification errors when a correlated two-factor population model was mis-specified as an uncorrelated two-factor model. Her findings revealed that the SEPC cutoff criterion can identify misspecification 70% of the overall replications in 80% of all the manipulated conditions in her study and it performed more accurately than the MI even when sample sizes and factor loading sizes were both small. Overall, there have been no consistent findings regarding the accuracy and stability of MI and/or EPC in detecting model misspecification; some studies revealed promising performance of MI and/or EPC [6, 36], but a preponderance of research found the performance less than acceptable [30, 32, 37, 38].

In summary, the majority of previous research only investigated misspecification in covariance structure and the findings are inconsistent. Hence, it is necessary to evaluate the effectiveness of using fit indices, MIs, and SEPC to detect misspecifications on the growth shape due to mislocations of turning points.

METHODS

Data Generation

A simulation study was conducted to address the above research questions. The population model used for data generation is a piece-to-piece linear growth model consisting of 7 equidistance time points and connected by one turning point. For simplicity, no covariates are included in the population model.

Based on the model defined by Equations (2–4), a total of 11 parameters are specified: four fixed effect coefficients (i.e., μ_{a1}, μ_{b1}, μ_{b2}, and γ) and seven variances and covariance of random effects (i.e., σ^2, $\tau_{\pi 00}$, $\tau_{\pi 10}$, $\tau_{\pi 20}$, $\tau_{\pi 11}$, $\tau_{\pi 21}$, $\tau_{\pi 22}$). **Table 1** presents the population parameter values specified based on previous studies (see e.g., [39]).

Design Factors

Based on previous findings regarding PGCM [5, 6, 27, 28], four design factors are considered, including (a) sample size, (b) the magnitude of change in the growth rate, (c) degree of severity in turning point misspecification, and (d) levels of non-normality.

TABLE 1 | Population parameters for the piecewise growth trajectory.

	Mean piecewise trajectory
a	2.5
b1	0.6
b2	0.54[a]
γ	3[b]
σ^2	1.0

$$\Phi$$

$$T_\pi = \begin{bmatrix} \tau_{\pi 00} & & \\ \tau_{\pi 10} & \tau_{\pi 11} & \\ \tau_{\pi 20} & \tau_{\pi 21} & \tau_{\pi 22} \end{bmatrix} = \begin{bmatrix} 0.200 & & \\ 0.050 & 0.100 & \\ 0 & 0.035 & 0.100 \end{bmatrix}$$

[a] Two levels of change in growth rate respectively at 0.54 and 0.44.
[b] Four levels of turning point are specified at 3, 3.5, 4, and 5.

Sample Size

The sample size was decided based on the empirical studies using piecewise latent growth curve modeling obtained from a literature search in PsycINFO (from 2010 to 2016). We chose three sample size conditions (75, 200, or 500 cases), representing approximately the minimum, 25th, and 50th percentiles of the sample size distribution.

Magnitude of Change in the Growth Rate

Based on Kwok et al.'s [6] study, we considered two levels in the magnitude change in growth rate: small change vs. medium change. Given that the growth rate in the first piece is 0.6, following Raudenbush and Liu's [39] effect size equation, the growth rate of the second piece is set to be 0.44 for the medium change condition and 0.54 for the small change condition.

Levels of Severity in Turning Point Misspecification

We generated data with four locations of turning point: 3, 3.5, 4, and 5 respectively. In the analysis model, the conventional PGCM specifies the turning point to be at time point 3. This is to mirror the two scenarios in reality: (1) the treatment began at time point 3, and was followed with an immediate change in growth rate (i.e., no misspecification); (2) the treatment effect was delayed (i.e., misspecification of 0.5, 1, or 2 time points).

Normality of Distributions

In longitudinal data, it is common to encounter non-normal data. To mimic real world data, we considered two conditions: normal and moderately skewed. For the moderately skewed distributions, the random effects were generated to have skewness of 1.5 and kurtosis of 6 respectively using Vale and Maurelli's [40] algorithm for simulating multivariate non-normal data. Such values are considered to be within the range of skewed distribution encountered in applied psychological research [41, 42].

In summary, the simulation used a 3 (number of sample size: 75 or 200 or 500) × 2 (magnitude of change in growth rate: small [B2 = 0.54] or medium [B2 =: 0.44]) × 4 (levels of severity of misspecification: 0, 0.5, 1, or 2 time points) × 2 (levels of distribution: normal or moderately skewed) factorial design to generate the data. A total of 500 replications were generated for each condition using SAS 9.4 Proc IML procedure [43], yielding 24,000 total data sets. Each replication was then fit with two different model specifications respectively: (1) the conventional PGCM with the turning point specified to be at the 3rd time point, and (2) the newly proposed PGCM with the turning point as an unknown parameter to be freely estimated. Both models were fit using Mplus version 7.4 [44] with Estimator = MLR. The Mplus code for the newly proposed PGCM is provided in the **Appendix** as a reference.

Analysis

Proper replications that reached convergence and had no improper solutions (e.g., negative variances) were retained for further analysis. The means and standard deviations of each fit index were presented along with their respective hit rates, which is a measure of the proportion of replications that successfully identified the correct or mis-specified models based on the recommended cutoff criteria (RMSEA $\leq$ 0.06; SRMR $\leq$ 0.08; TLI $\geq$ 0.95; FCI $\geq$ 0.95) recommended by Hu and Bentler [22].

For Modification Indexes (MI), because the purpose is to detect growth shape misspecification due to the incorrectly located turning point, we restricted the search of MIs among the loadings of time points 3 to 7 associated with the 1st piece and the 2nd piece growth factors. To maintain the family-wise Type I error at the 0.05 level, we adjusted the alpha level to be at 0.005 because there are a total number of 10 potential fixed loadings to be modified. Therefore, the threshold of a MI to be considered significant was 7.88 ($df = 1$ and $\alpha = 0.005$). For SEPCs (the fully standardized Expected Parameter Changes), we used the cutoff value of 0.2 as recommended by Saris et al. [35].

Estimates of the turning point, growth parameters, their corresponding standard errors and the random effects were summarized across all proper replications for each condition. The standardized biases of the estimates [i.e., $B(\hat{\theta}) = (\hat{\theta} - \theta)/S(\hat{\theta})]^1$ were calculated. The mean of the standardized bias is equivalent to a Cohen's d, which measures the standardized distance between the estimate and the parameter. Based on the guidelines for Cohen's d, the value of less than 0.14 is considered acceptable. For turning point estimates, the unstandardized biases [i.e., $B(\hat{\theta}) = \hat{\theta} - \theta$] were also calculated to show the bias in the original metric of time.

Analysis of variance (ANOVA) was then used to examine the impact of the design factors on the bias of the parameter estimates. The eta-squared ($\eta^2 = SS_{\text{Effect}}/SS_{\text{Total}}$) effect size was computed and reported as a measure of practical significance. Effects were considered substantial with the eta-squared greater than 0.1.

RESULTS

Model convergence was explicitly examined to ensure a clear and appropriate analysis of the results. All 500 replications in each of the designed conditions, estimated using the conventional

[1] Where $\hat{\theta}$ is the parameter estimate, θ the population parameter value, and $S(\hat{\theta})$ the standard deviation of the estimates across 500 replications.

PGCM with the turning point determined a priori, converged successfully with no improper solutions. For the PGCM with an unknown turning point, the average convergence rate was around 80%. Non-convergence or improper solutions occurred more often with smaller sample size. The average convergence rate with no improper solutions for replications estimated using the PGCM with unknown turning points is 68% ($n = 75$), 81% ($n = 200$), and 88% ($n = 500$) across all other designed conditions.

Performance of Fit Indices under the True Models

The means and standard deviations (SDs) of the examined fit indices (i.e., Chi-square test statistics, CFI, TLI, SRMR, and RMSEA) for the conventional PGCM and the newly proposed PGCM were summarized across all proper replications under data distributions (see **Table 2**). For the conventional PGCM with the turning point correctly specified a priori, when distributions were normal, the mean of $\chi^2_{(df = 19)}$ was 19.59 and the SD was 6.35. The values were similar to the mean of χ^2 (19.78) and the SD (6.42) for the same model specification when data distributions were moderately skewed. For the newly proposed PGCM, the mean of $\chi^2_{(df = 18)}$ was 19.16 and the SD was 12.64 for normal distributions, and was 22.35 and 18.14 for moderately skewed data distributions. Type I error rates associated with the Chi-square test for the conventional PGCM (i.e., the rate of rejecting a correctly specified model) were almost identical for normal (i.e., 6.83%) and skewed distributions (i.e., 6.60%).

For the newly proposed PGCM, the Type I error rate was 7.27% when distributions were normal, which was lower than that for the skewed distributions (Type I error rate = 10.67%). This suggests that the newly proposed PGCM could be sensitive to data distributions, and deviation from normality could result in higher rejection rate even when the model was appropriately specified.

Table 2 also presented the means and SDs of RMSEA, CFI, TLI, and SRMR, as well as their hit rates, which are the percentages of replications that correctly identified the true models. The means of RMSEA were below 0.06 across the conditions. The hit rates of the indexes were 94.6 and 92% respectively in both model specifications for normal distributions, and dropped to 90 and 85% when distributions were moderately skewed. Contrarily, the hit rates of SRMR were 86.4 and 90.2% in both models for normal distributions but increased substantially to almost 100% for conventional PGCM and 98.2% for newly proposed PGCM when data distributions were moderately skewed. CFI and TLI had means of 1.0 and almost 100% in hit rates across all the conditions in the correctly specified model.

The means and SDs of the modification indices[2] (MI) as well as the percentage of the replications that had significant MI value associated with the targeted fixed parameters were presented in **Table 3**. Additional information summarized in **Table 3** includes the range of SEPCs (the fully standardized Expected Parameter Changes) and the percentage of SEPCs larger than 0.2. The mean of MI ranged from 5.34 to 5.89 and the percentage of significant MI (≥7.88) ranged from 2.4 to 3.4% across the design factors. The range of SEPCs became narrower with the increase of sample size regardless of distributions and the change in growth rate. The percentages of SEPCs larger than 0.2 ranged from 0 to 4%.

Performance of Fit Indexes in PGCMs with Mis-Specified Turning Points

Table 4 summarized the descriptive information for the χ^2 test statistic, RMSEA, SRMR, CFI, and TLI across all proper replications in conventional PGCMs when the turning point is mis-specified.

When distributions were normal, the means and SDs of the χ^2 test statistic increased from 20.56 and 6.79 to 27.00 and 10.10 as the degree of turning point misspecification

[2] MI is not available in Mplus for the PGCM with unknown turning points due to nonlinear constraints in fitting the model.

TABLE 2 | Descriptive statistics of fit indices when the true turning point was at time point 3.

Fit indices	Conventional PGCM with turning point correctly specified at 3				PGCM with unknown turning point estimated based on data			
	Normal longitudinal data		Skewed longitudinal data		Normal longitudinal data		Skewed longitudinal data	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Chi-square	19.59	6.35	19.78	6.42	19.16	12.64	22.35	18.14
	(Type I error rate = 6.83%)		(Type I error rate = 6.60%)		(Type I error rate =7.27%)		(Type I error rate =10.67%)	
RMSEA	0.02	0.03	0.02	0.03	0.02	0.03	0.03	0.04
	(Hit rate = 94.6%)		(Hit rate = 90%)		(Hit rate = 92%)		(Hit rate = 85%)	
CFI	1.00	0.01	1.00	0.00	1.00	0.03	1.00	0.01
	(Hit rate = 100%)		(Hit rate = 100%)		(Hit rate = 100%)		(Hit rate = 100%)	
TLI	1.00	0.02	1.00	0.01	1.00	0.04	1.00	0.01
	(Hit rate = 98.4%)		(Hit rate = 100%)		(Hit rate = 100%)		(Hit rate = 98%)	
SRMR	0.05	0.03	0.02	0.01	0.05	0.02	0.03	0.02
	(Hit rate = 86.4%)		(Hit rate = 99.8%)		(Hit rate = 90.2%)		(Hit rate = 98.2%)	

Degrees of freedom for the conventional PGCM was 19. Degrees of freedom for PGCM with unknown turning points was 18.

TABLE 3 | Descriptive statistics of modification indices (MI) for the conventional PGCM with the turning point correctly specified to be at 3.

Impact factors		Normal distribution					Skewed distribution				
Change rate	Sample size	MI			SEPC		MI			SEPC	
		Mean	SD	Percent of MI (≥7.88)	(Min, Max)	Percent of SEPC (≥0.2)	Mean	SD	Percent of MI (≥7.88)	(Min, Max)	Percent of SEPC (≥0.2)
Small	75	5.47	1.67	2.4	(−0.31, 0.24)	4	5.31	1.66	2.2	(−0.18, 0.30)	1
	200	5.89	2.04	2.2	(−0.16, 0.16)	0	5.80	2.49	2.2	(−0.15, 0.13)	0
	500	5.34	1.40	1.4	(−0.08, 0.09)	0	5.60	1.95	2.4	(−0.07, 0.102)	0
Medium	75	5.70	1.81	2.6	(−0.28, 0.31)	4	5.53	1.65	2.4	(−0.13, 0.11)	0
	200	5.55	1.52	2.0	(−0.14, 0.15)	0	5.55	1.81	3.4	(−0.18, 0.26)	0
	500	5.58	1.63	2.0	(−0.10, 0.09)	0	5.67	1.79	2.6	(−0.07, 0.07)	0

TABLE 4 | Descriptive statistics of the chi-square test statistic and fit indices of the conventional PGCM with the turning point mis-specified to be at 3.

Distribution	True turning point	Chi-square		RMSEA		CFI		TLI		SRMR	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Normal	3.5	20.56	6.79	0.02	0.03	1.00	0.01	1.00	0.01	0.05	0.02
		(Power = 0.08)		(Hit rate = 9.4%)		(Hit rate = 0.00%)		(Hit rate = 1.26%)		(Hit rate = 12.6%)	
	4	23.67	8.13	0.03	0.03	1.00	0.01	1.00	0.02	0.05	0.02
		(Power = 0.19)		(Hit rate = 13.6%)		(Hit rate = 0.01%)		(Hit rate = 1.50%)		(Hit rate = 13.7%)	
	5	27.00	10.10	0.03	0.03	0.99	0.01	0.99	0.01	0.05	0.02
		(Power = 0.33)		(Hit rate = 17.5%)		(Hit rate = 0.04%)		(Hit rate = 1.56%)		(Hit rate = 15.0%)	
Skewed	3.5	37.39	15.96	0.06	0.03	0.99	0.01	0.99	0.01	0.03	0.01
		(Power = 0.60)		(Hit rate = 48.5%)		(Hit rate = 0.05%)		(Hit rate = 0.67%)		(Hit rate = 0.27%)	
	4	81.46	45.60	0.11	0.03	0.98	0.01	0.97	0.01	0.04	0.01
		(Power = 0.92)		(Hit rate = 96.9%)		(Hit rate = 1.63%)		(Hit rate = 5.57%)		(Hit rate = 1.07%)	
	5	92.39	53.68	0.12	0.03	0.98	0.01	0.97	0.01	0.05	0.01
		(Power = 0.94)		(Hit rate = 98%)		(Hit rate = 2.00%)		(Hit rate = 6.00%)		(Hit rate = 2.00%)	

increased from 0.5 to 2 time points. With skewed distributions, the changes were much greater, from 37.39 and 15.96 in mean and SD to 92.39 and 53.68 as the degree of turning point misspecification increased from 0.5 to 2 time points. The empirical power to detect turning point misspecification for normally distributed data was low (power = 0.33 with 2 time points misspecification). However, when distributions were moderately skewed, the empirical power to detect model misspecification reached 0.92 with 1 time point misspecification and 0.94 with 2 time point misspecification. It is suggestive that the misspecification of turning point was confounded with deviations from multivariate normality. The χ^2 test statistic detects the non-normality in the distribution, not necessarily the turning point misspecification.

As shown in **Table 4**, when distributions were normal, the means of RMSEA were below 0.06 (the cutoff criteria). The hit rates were small, ranging from 9.4, 13.6, to 17.5%, showing low sensitivity to turning point misspecification. However, when it came to skewed distributions, the means of RMSEA increased from 0.06 to 0.12 as the severity of misspecification increased from 0.5 to 2 time points. The same increase trend was observed for hit rates, increasing from 48.5% with 0.5 time point misspecification to 96.9 and 98% when the misspecification was by 1 or 2 time points respectively.

The means of CFI and TLI showed almost no deviation from 1.0 with very small SDs across all conditions. The hit rates of both CFI and TLI were close to 0 when distributions were normal and increased to about 6.00% when distributions are skewed, regardless of the increased severity in turning point misspecification. The performance of CFI and TLI was the least desirable in capturing the misspecification in turning point.

The means and SDs of SRMR remained the same (0.05, smaller than the cutoff value 0.08) across the different levels of severity in turning point misspecification with normal distributions. An increasing trend was observed in the hit rates with the increase of misspecification severity, but not to the extent of being effective in detecting the model misspecification. The hit rates in skewed data conditions were much smaller in size than the values in the conditions of normal distributions.

Table 5 summarized the performance of MI and SEPC in identifying the misspecification in conventional PGCM when the a priori turning point was mis-specified. Holding the severity of misspecification constant, the means of MI were found to increase with the increase of the sample size and with the change from normal distributions to skewed distributions. With normal distributions, the percentage of replications with MI exceeding the threshold of 7.88 ranged between 2.7% (0.5 time point misspecification with a sample size of 75) and 40.0% (2 time points misspecification with a sample size of 500). The percentage increased substantially when the distributions were moderately skewed.

SEPC showed a pattern of increasingly narrower range with the increase of sample size after keeping the levels of severity in turning point misspecification constant. The percentages of replications with SEPC exceeding the threshold of 0.2 were below 5% across all conditions. Overall, SEPC based on MI was a poor indicator in detecting the mis-specified locations of turning points.

Standardized Bias of Fixed Effect Estimates

Table 6 presents the means of the standardized bias of fixed effect estimates for the correctly specified conventional PGCMs (i.e., the turning point was specified correctly a priori) and PGCMs with an unknown turning point estimated based on data when the true turning point is 3 in the population. On average, with conventional PGCMs, the standardized bias of fixed effect estimates of the Intercept (a), 1st Slope (b_1) and 2nd Slope (b_2) ranged from 0 to 0.08, negligibly small across the design factors. For PGCMs with unknown turning points, the standardized bias of fixed effect estimates a, b_1, and b_2 ranged from 0.00 to 0.32. Larger standardized biases were found under the normal distribution condition than the skewed distribution. This is counterintuitive; however, a closer examination revealed that the unstandardized biases were larger under the skewed distribution condition. The standardized biases looked smaller, because the standard deviations of the estimates were inflated under the skewed distribution.

The interaction between data distributions and change in growth rate explained a substantial amount of variation in the biases of the estimate of a ($\eta^2 = 0.15$) (see **Figure 1A**) and of b_1 ($\eta^2 = 0.16$) (see **Figure 1B**). The interaction effects between data distributions and sample size ($\eta^2 = 0.13$) (see **Figure 2A**) and between sample size and change rate ($\eta^2 = 0.12$) (see **Figure 2B**) were found to account significantly for the variations of the biases of the estimates of $b_{2.}$

As summarized in **Table 7**, when the conventional PGCM was mis-specified due to the mislocation of the turning point, with normal distributions and a small change in growth rate, the biases of the estimate of *a* were acceptable regardless

TABLE 5 | Descriptive statistics of modification indices (MI) for the conventional PGCM with the turning point mis-specified to be at 3.

Impact factors		Normal distribution					Skewed distribution				
True turning point	Sample size	MI			SEPC		MI			SEPC	
		Mean	SD	Percent of MI (≥7.88)	(Min, Max)	Percent of SEPC (≥0.2)	Mean	SD	Percent of MI (≥7.88)	(Min, Max)	Percent of SEPC (≥0.2)
3.5	75	5.66	1.76	2.7	(−0.33, 0.22)	3	5.99	2.35	6.8	(−0.37, 0.26)	1.60
	200	5.85	2.32	3.1	(−0.24, 0.18)	0	6.52	2.89	15.1	(−0.18, 0.17)	0.00
	500	6.11	2.21	4.4	(−0.12, 0.09)	0	7.42	3.58	34.8	(−0.12, 0.12)	0.00
4	75	5.78	1.92	3.0	(−0.37, 0.24)	4	7.36	3.88	20.4	(−0.32, 0.35)	4.40
	200	6.25	2.49	5.9	(−0.23, 0.18)	0	9.07	5.47	46.7	(−0.26, 0.16)	0.37
	500	7.31	3.43	15.3	(−0.13, 0.12)	0	14.00	9.45	90.7	(−0.19, 0.14)	0.00
5	75	6.07	3.13	4.7	(−0.28, 0.18)	2	8.29	4.17	33.3	(−0.29, 0.22)	2.22
	200	6.87	3.02	15.2	(−0.21, 0.14)	0	11.92	7.26	72.5	(−0.24, 0.14)	0.13
	500	9.17	5.01	40.0	(−0.15, 0.09)	0	20.91	14.35	99.2	(−0.18, 0.12)	0.00

TABLE 6 | Mean standardized biases of fixed effects when the true turning point was at time point 3.

Impact factors		Conventional PGCM			PGCM with an unknown turning point		
Distribution	Growth rate change	Correct specification of true turning point at 3			Estimated based on data		
		Intercept (a)	1st Slope (b_1)	2nd Slope (b_2)	Intercept (a)	1st Slope (b_1)	2nd Slope (b_2)
Normal	0.54 (Small)	0.08	−0.03	−0.03	0.00	0.18	−0.16
	0.44 (Medium)	−0.02	0.02	0.02	−0.2	0.32	−0.12
Skewed	0.54 (Small)	0.03	0.03	−0.03	0.03	0.01	−0.06
	0.44 (Medium)	0.02	−0.04	0.00	0.01	−0.07	−0.06

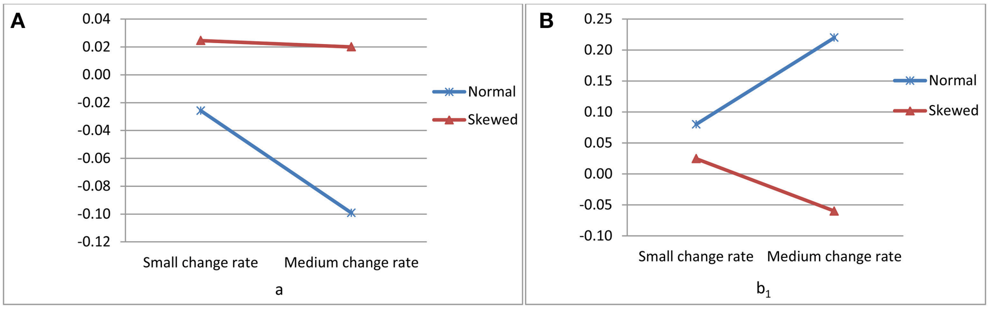


FIGURE 1 | (A) Interaction effect between change in growth rate and data distributions on the standardized bias of the intercept (a) estimate and **(B)** Interaction effect between change in growth rate and data distributions on the bias of b_1 estimate.

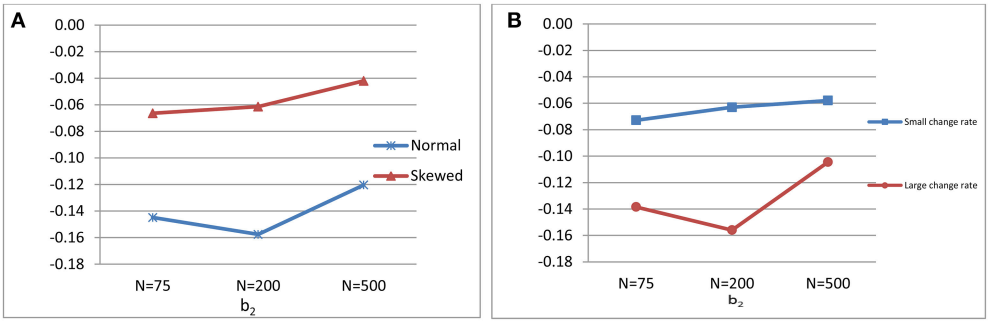


FIGURE 2 | (A) Interaction effect between sample size and data distributions and **(B)** Interaction effect between sample size and change in growth rate on the standardized bias of b_2 estimate.

TABLE 7 | Mean standardized biases of fixed effects when the true turning point was at 3.5, 4, or 5.

Impact factors		**Conventional PGCM**			**PGCM with an unknown turning point Estimated based on data**		
Distribution	**Growth rate change**	**Intercept (a)**	**1st Slope (b_1)**	**2nd Slope (b_2)**	**Intercept (a)**	**1st Slope (b_1)**	**2nd Slope (b_2)**
		Mis-specified the true turning point at 3.5 to be 3			**The true turning point at 3.5**		
Normal	0.54 (Small)	−0.03	0.15	0.23	−0.05	0.14	−0.08
	0.44 (Medium)	−0.05	0.31	0.66	0	0.1	−0.14
Skewed	0.54 (Small)	0.01	0.08	0.14	0.02	−0.06	−0.04
	0.44 (Medium)	−0.08	0.16	0.33	−0.01	−0.06	−0.07
		Mis-specified the true turning point at 4 to be 3			**The true turning point at 4**		
Normal	0.54 (Small)	−0.06	0.22	0.48	−0.03	0.12	−0.07
	0.44 (Medium)	−0.16	0.64	1.3	−0.11	0.24	−0.2
Skewed	0.54 (Small)	−0.03	0.12	0.3	0.02	0.08	−0.01
	0.44 (Medium)	−0.02	0.21	0.67	0.03	−0.04	−0.16
		Mis-specified the true turning point at 5 to be 3			**The true turning point at 5**		
Normal	0.54 (Small)	−0.03	0.22	1.19	−0.01	0.09	−0.09
	0.44 (Medium)	−0.19	0.74	3.16	−0.1	0.19	−0.27
Skewed	0.54 (Small)	−0.03	0.07	0.6	0.02	0	−0.01
	0.44 (Medium)	−0.03	0.28	1.59	0.05	0.06	−0.05

of the misspecification severity. However, the increase in misspecification severity from 0.5 to 2 time points led to an increase from 0.15 to 0.22 in the mean standardized bias for b_1 and from 0.23 to 1.19 for b_2. The biases were larger when the change in growth rate was medium, resulting in an increase in the bias from 0.31 to 0.77 for b_1 and from 0.66 to 3.16 for b_2, with increase in misspecification severity from 0.5 to 2 time points.

When distributions were skewed, the corresponding bias was mitigated to some degree but was still considered unacceptable particularly when the change in growth rate was medium. For example, the mean bias for b_2 increased from 0.33 to 1.59 with increase in misspecification severity from 0.5 to 2 time points. However, when using PGCM with unknown turning points, biases were considered small (about 0.20) for almost all fix effect estimates across almost all conditions in spite of the different distributions in the data.

ANOVA was used to partition the total variance in the standardized biases associated with the effects of the six design factors. **Table 8** presented the eta-squared and the statistical significance of the main effects and interactions of the design factors on standardized bias. Statistically significant effects ($p < 0.05$) were marked with asterisks and were bolded if they were found to be practically significant ($\eta^2 > 0.1$). The design factors of data distributions ($\eta^2 = 0.17$) and change rate in growth ($\eta^2 = 0.10$) had substantial main effects on the bias of the mean of intercept (α). Data distribution was found to have substantial main effect ($\eta^2 = 0.20$), model specification and the level of severity in turning point misspecification were found to have significant interaction effects respectively on the bias associated with the estimates of b_1 ($\eta^2 = 0.13$) and with the mean of b_2 ($\eta^2 = 0.17$), as **Figure 3A** shows.

Standardized Bias of Variance-Covariance Estimates

Table 9 shows the means of the standardized bias of variance components estimates broken down by model specification, distributions and sample size, the factors consistently found to be systematically related to the observed bias of the estimates of the variance components. When the fitting model was PGCM with unknown turning points, for normal data distributions, the mean biases of the estimated variance components were negligibly small, ranging from −0.03 to 0.15. On the other hand, when the fitted model was the conventional PGCM

TABLE 8 | Effect sizes of the impacts of the design factors on the standardized bias of estimated model parameters.

Impact factors	Fixed effect estimates				Estimates of standard error of fixed effects				Estimates of variance components					
	a	b1	b2	γ	a	b1	b2	γ	Variance			Covariance		
									$\tau_{\pi 00}$	$\tau_{\pi 11}$	$\tau_{\pi 22}$	$\tau_{\pi 10}$	$\tau_{\pi 20}$	$\tau_{\pi 21}$
Model specification	0.01	0.07*	**0.26***		0.02	0.05*	0.02		0.00*	0.00*	0.01*	0.00	**0.28***	**0.11***
Data distribution	**0.17***	**0.20***	0.01*	**0.52***	0.00	0.04*	**0.10***	**0.72***	**0.77***	**0.75***	**0.63***	**0.72***	0.07*	**0.53***
Turning point	0.04	0.09*	**0.17***	**0.23***	0.01	0.06	**0.14***	**0.**09*	0.00	0.01*	0.01*	0.02*	**0.12***	0.01*
Change rate	**0.10***	0.07*	0.03*	0.00	0.01	0.00	0.00	0.00	0.11	0.00	0.00	0.00	0.00	0.00
Sample size	0.01	0.03*	0.03*	0.00	0.02	0.03	0.01	0.03*	0.00*	**0.12***	0.08*	**0.12***	0.04*	**0.11***
Model specification × Data distribution	0.02	0.00	0.03*		0.07*	0.09*	0.04*		0.00*	0.00	**0.11***	0.00	**0.16***	0.08*
Model specification × Turning point	0.08*	**0.13***	**0.17***		0.01	0.06	0.04		0.00	0.00*	0.00*	0.00*	**0.11***	0.02*
Model specification × Change rate	0.00	0.03*	0.05*		0.00	0.01	0.00		0.00	0.00	0.00	0.00	0.00	0.00
Model specification × Sample size	0.01	0.04*	0.03*		0.02	0.00	0.05*		0.00*	0.00*	0.00	0.00	0.02*	0.01*
Data distribution × Turning point	0.03	0.00	0.02	0.05*	0.05	0.07*	0.00	**0.14***	0.00*	0.00*	0.00	0.02*	0.08*	0.00
Data distribution × Change rate	0.05*	0.04*	0.00	0.00	0.00	0.00	0.01	0.00	0.00*	0.00	0.00	0.00	0.00	0.00
Data distribution × Sample size	0.00	0.00	0.00	0.03*	**0.12***	0.03	0.05*	0.00	0.11	**0.10***	**0.13***	**0.10***	0.02*	**0.10***
Turning point × Change rate	0.02	0.03*	0.03*	0.01	0.06	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Turning point × Sample size	0.08*	0.03	0.02	**0.12***	0.06	0.06	0.07	0.00	0.00	0.00	0.00	0.01*	0.01	0.00
Change rate× Sample size	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Factors with a statistically significant effect (p < 0.05) were marked with an asterisk and was bolded if they were found to have practically significant effect (eta squared > 0.1).

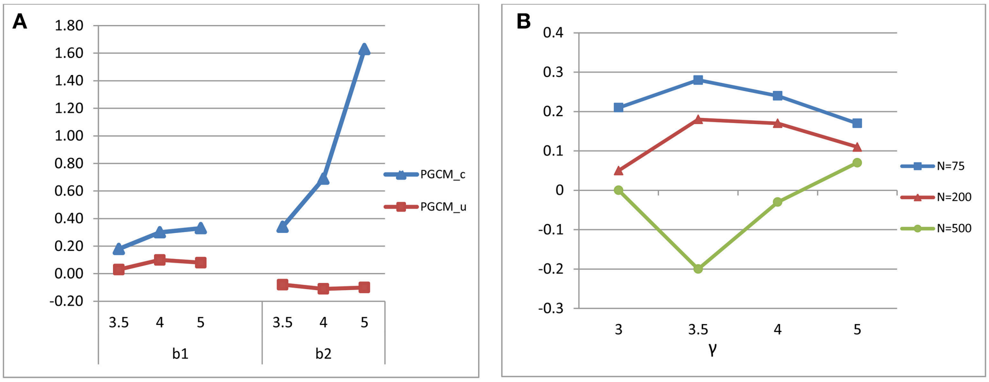


FIGURE 3 | (A) Interaction effect between the locations of true turning point and model specification (Model Specification × Turning point) on the standardized bias of fixed effect estimate of b_1 and b_2. **(B)** Interaction effect between sample size and the locations of true turning point (Turning point × Sample size) on the standardized bias of turning point (γ) estimate in PGCM with unknown turning points.

TABLE 9 | Mean standardized biases of variance components estimates.

Impact factors			**Standardized bias of variance and covariance of growth factors**					
Model specification	**Distribution**	**Sample size**	$\tau_{\pi 00}$	$\tau_{\pi 11}$	$\tau_{\pi 22}$	$\tau_{\pi 10}$	$\tau_{\pi 20}$	$\tau_{\pi 21}$
PGCM_c	Normal	75	−0.04	0.14	−0.65	0.01	0.05	0.04
		200	−0.05	0.31	−0.96	0.02	0.09	0.05
		500	−0.03	0.51	−1.52	0.07	0.14	0.09
	Skew	75	1.67	2.60	2.73	1.15	0.37	0.90
		200	2.77	4.25	4.58	1.88	0.61	1.49
		500	4.42	6.63	7.20	3.03	1.00	2.31
PGCM_u	Normal	75	−0.01	0.06	0.11	−0.04	−0.03	−0.02
		200	−0.06	0.11	0.14	0.01	−0.02	−0.03
		500	−0.06	0.11	0.15	0.08	−0.02	−0.03
	Skew	75	1.73	2.42	1.03	1.13	−0.16	0.25
		200	2.94	3.92	1.99	1.84	−0.19	0.52
		500	4.59	6.10	4.69	3.03	−0.09	1.18

PGCM_c is conventional PGCM with turning point specified a priori; PGCM_u is PGCM with unknown turning points.

with mis-specified turning points, large biases were observed in the variance components estimates. The biases increased as sample size increased. For example, an increase of sample size from 75 to 500 led to an increase in the mean biases from 0.14 to 0.51 for $\tau_{\pi 11}$ and from −0.65 to −1.52 for $\tau_{\pi 22}$. In addition, when the data distributions were skewed, the variance components were highly biased for both models and in general the biases increased with the increase of sample size.

Bias of the Turning Point Estimates

The accuracy of the turning point estimates was examined using both the standardized and the unstandardized bias due to the practical meaning of the metric of time points. As summarized in **Table 10**, the maximum mean unstandardized bias was around 0.25 when sample size is 75, growth rate change is small, and distribution is normal, indicating that the estimated turning point is 0.25 time point away from the true turning point under those conditions. The minimum mean unstandardized bias was around 0.01 when sample size is 500, growth rate change is medium, and distribution is normal. Regardless of the locations of the true turning point, the mean unstandardized biases decreased with the increase in sample size and the increase in the change of growth rate (from small to medium). Though similar trends were observed with skewed distributions, the values of the unstandardized biases were much larger.

On the other hand, taking the variation in the turning point estimates into consideration, the standardized bias was much smaller under the normal distribution condition ($\eta^2 = 0.52$). An interaction effect was found between the location of the turning point and sample size ($\eta^2 = 0.12$). As shown in **Figure 3B**, the standardized biases were the smallest when the turning point was located at time point 4 and when the sample size was moderately large ($N = 500$).

TABLE 10 | Mean unstandardized and standardized biases of turning point (γ) estimates.

Impact factors			**The location of the turning point**			
Distribution	**Growth rate change**	**Sample size**	***t* = 3**	***t* = 3.5**	***t* = 4**	***t* = 5**
Normal	0.54 (Small)	75	0.25 (0.29)	0.04 (0.05)	0.07 (0.09)	−0.26 (−0.32)
		200	0.00 (0.00)	0.14 (0.19)	−0.02 (−0.04)	−0.12 (−0.18)
		500	0.02 (0.04)	0.02 (0.04)	−0.05 (−0.10)	−0.03 (−0.07)
	0.44 (Medium)	75	0.04 (0.06)	0.07 (0.09)	0.01 (0.02)	−0.20 (−0.26)
		200	−0.03 (−0.06)	0.03 (0.05)	0.03 (0.04)	−0.02 (−0.04)
		500	−0.04 (−0.09)	0.06 (0.17)	−0.00 (−0.01)	0.03 (0.07)
Skewed	0.54 (Small)	75	0.23 (0.33)	0.40 (0.56)	0.24 (0.44)	−0.04 (−0.10)
		200	0.15 (0.24)	0.28 (0.42)	0.16 (0.37)	0.02 (0.08)
		500	0.04 (0.13)	0.12 (0.28)	0.06(0.26)	0.02 (0.12)
	0.44 (Medium)	75	0.27 (0.36)	0.38 (0.52)	0.22 (0.41)	−0.02 (−0.07)
		200	0.16 (0.28)	0.28 (0.44)	0.19 (0.42)	0.00 (0.01)
		500	0.03 (0.12)	0.08 (0.22)	0.10 (0.33)	0.01 (0.10)

The values outside the parentheses are unstandardized bias. The values within the parentheses are standardized bias.

Standardized Bias of Standard Errors of Fixed Effects and Turning Point Estimates for PGCM with Unknown Turning Point

Table 11 showed the means of the standardized bias of the SEs of fixed effect and turning point estimates for PGCM with unknown turning point. The mean standardized bias of SEs of a, b_1 and b_2 were negligibly small, ranging from 0.00 to 0.07 in absolute values. However, large biases were found in the SEs associated with the estimates of the turning point (γ). When the distributions were normal, the observed bias of the SEs of γ were acceptable, ranging from −0.03 to 0.26; when the distributions were skewed, the estimates of the SEs of γ were highly biased and underestimated with exception to the condition when the turning point was located at the 5th time point.

As observed in **Table 8**, data distributions and sample size had statistically and practically significant interaction effect ($\eta^2 = 0.12$) on the observed bias of SEs of a. The SEs of b_2 were found to be substantially affected by data distribution ($\eta^2 = 0.10$) and the severity of turning point misspecification ($\eta^2 = 0.14$). Data distribution and the locations of the true turning point exhibited a significant interaction effect ($\eta^2 = 0.14$) on the bias of the SEs of the turning point (γ). The earlier the turning point is located at the time series, the larger the biases are associated with the estimates of the SEs of the turning point (γ), and such biases are even larger with moderately skewed data distributions.

DISCUSSION

The study investigated the impacts of mis-specified turning point on growth trait estimation in conventional PGCMs that require turning points to be specified a priori. We examined the sensitivity of generally used model fit diagnostics [i.e., χ^2 test statistic, RMSEA, CFI, TLI, SRMR, modification index (MI), and SEPC] in detecting specification errors in conventional PGCMs due to turning points mislocation. In addition, the performance of an alternative procedure, PGCMs with unknown turning points (i.e., the turning point is treated as a parameter to be estimated based on data) was evaluated. The design factors considered in the simulation study were locations of true turning point (respectively at time point 3, 3.5, 4, or 5), sample size (75 or 200 or 500), and data distributions (normal vs. moderately skewed). This section summarized and discussed the results of the study.

Impact of Turning Point Misspecification

Misspecification of the turning point in conventional PGCM was found to have a substantial impact on the fixed effects estimates of 1st Slope (b_1) and 2nd Slope (b_2). The biases were considered acceptable only when the turning point was mis-specified by 0.5 time point with a small change in growth rate between the 1st and 2nd piece. Misspecification of a turning point earlier than its true location would result in overestimation of the growth rates in b_1 and b_2. Overall, the more severe the misspecification of the turning point is, the greater the impact is on the estimates, and the more misrepresented the growth trait estimates are for the population data. Such consequences are exacerbated when the change in growth rate is medium.

As expected, misspecification of the turning point also gives rise to unacceptably large biases with respect to the estimated variance components. The variances of the slopes of the 1st and 2nd piece are underestimated, with the latter being more severely underestimated. Since the variance of the slope factors reflects inter-individual differences in growth rates, the underestimation results may lead to the wrong conclusion that individuals have similar growth process. For applied researchers who are interested in using individual level predictors to predict the variation in growth rates, the deflated variance component estimates may attenuate the relationship between the predictors and growth rates, leading to misleading inferential conclusions.

TABLE 11 | Mean standardized biases of standard errors of fixed effects and turning point estimates.

Impact factors		**Normal distribution**				**Skewed distribution**			
Location of turning point	**Sample size**	**Intercept (a)**	**1st Slope (b_1)**	**2nd Slope (b_2)**	**γ**	**Intercept (a)**	**1st Slope (b_1)**	**2nd Slope (b_2)**	**γ**
3	75	0.01	−0.02	0.03	0.18	0.00	−0.03	−0.05	−0.60
	200	0.01	0.11	0.02	0.26	−0.03	−0.01	0.00	−0.72
	500	0.01	0.04	−0.02	0.05	0.02	0.00	−0.03	−0.65
3.5	75	0.01	−0.02	0.07	0.19	−0.02	−0.01	0.03	−0.57
	200	0.02	0.00	0.02	0.05	−0.04	−0.03	−0.02	−0.72
	500	−0.01	−0.02	0.05	0.06	0.00	0.00	0.01	−0.77
4	75	0.03	0.02	0.00	0.22	−0.02	−0.01	−0.04	−0.47
	200	−0.01	0.03	0.02	0.04	0.03	0.00	0.00	−0.63
	500	−0.04	−0.02	−0.02	0.00	0.02	0.00	−0.02	−0.65
5	75	0.02	0.00	0.06	0.17	0.00	0.02	−0.01	−0.06
	200	−0.02	0.01	0.06	0.04	−0.02	0.03	0.02	−0.12
	500	−0.02	0.04	0.01	−0.03	0.03	0.01	0.03	−0.11

Sensitivity of Model Fit Index Diagnostics

An optimal identification of the location of a turning point a priori is important for conventional PGCMs. When the location of a turning point was mis-specified in conventional PGCMs, our simulation results indicated that the model fit indices [i.e., χ^2 test statistic, RMSEA, CFI, TLI, SRMR, modification index (MI), and SEPC], generally did not perform effectively in detecting the misspecification errors. The performance of the overall model χ^2 test was not only affected by the severity of misspecification in turning point but also by the incidental characteristics of the data (e.g., data distributions, sample size). The magnitude of χ^2 test statistic increased as the distribution changes from normal to skewed and with the increase in sample size. Such undesirable characteristics of χ^2 test statistic were already confirmed in many studies (see e.g., [35, 45–47]). Additionally, χ^2 test statistic showed a lack of adequate power to detect model misspecification in almost all conditions with exception to conditions where the data distributions were moderately skewed and severity in misspecification was by 1 time point or more.

As a function of χ^2, RMSEA performed similarly as χ^2 test statistic. RMSEA was not sensitive to the degree of misspecification in turning point when distributions were normal. Similar to χ^2 test statistic, it was found to be relatively more effective only when distributions were moderately skewed and the turning point was mis-specified by 1 or more time points. However, such seemingly high power of RMSEA in non-normal distributions should be taken with caution, as warned by Nevitt and Hancock [48], the apparent advantage of high power of RMSEA is a result of the inflated χ^2 test statistic when multivariate normality is violated. Nor were CFI, TLI, and SRMR effective in capturing the misspecifications under any of the design conditions. Although previous studies showed that the three fit indices are effective in detecting the specification error when a piecewise growth trajectory is mis-specified as linear (see [27, 28]), our study shows that the three fit indices do *not* work well when the misspecification is on the location of the turning point rather than the linearity of the trajectory.

The findings with regards to the performance of the MI and the SEPC showed that MI tended to be more accurate in skewed distributions particularly in conditions where the severity in turning point misspecification was by at least 1 time point and the sample size was moderately large ($N = 500$). It is not surprising that the performance of modification indices are also influenced by data distributions; modification index is a function of χ^2 test, basically a univariate delta χ^2 tests computed on each fixed parameter if freely estimated. Contrary to the recommendation of Saris et al. [35] that an SEPC $\geq$ 0.2 indicates a substantial misspecification, we found that SEPC was a poor indicator of turning point misspecification in PGCM.

Performance of PGCM with Unknown Turning Points

Overall, the PGCM with unknown turning points was found to perform very well in recovering the fixed effects and the random effects when the longitudinal responses follow a multivariate normal distribution. However, when data distributions deviate from normality, relatively large biases were found to be associated with the standard error estimates of the turning point and the random effects of growth factors. The biases of the turning point estimates are small to moderate in general. It is interesting that when the location of the turning point is at time point 4, the estimation of the turning point becomes highly accurate regardless of data distribution types. This finding to some degree corresponds to the results in Kohli and Harring's [5] study which showed that the locations of the turning point were systematically related to the relative bias of growth parameters, particularly with the estimation of the mean of the slope of the 2nd piece. Specifically, the earlier the turning point is located in the time series, the larger the bias is. However, their study differed from ours in two major aspects: their population model was a second-order piecewise latent growth curve model; though they evaluated the model performance in the recovery of growth parameters with regards to the estimation of the intercept, slopes of the 1st and 2nd piece, the focus was not on the estimation of the turning point and therefore, no relevant findings were discussed in their study.

Recommendations

In longitudinal data analysis, if a turning point is hypothesized at a specific time point, applied researchers tend to specify an a priori piecewise linear model to capture the turning point and examine whether the model fits the data. The simplest approach to evaluate model fit is through the use of model fit indices and MIs, however, the present findings showed that those generally used model fit diagnostics are not accurate or effective in detecting specification errors related to the turning point location. Unless the misspecification is severe (i.e., by at least 1 time point) and the longitudinal data follow a multivariate moderately skewed distribution, the generally used fit indices are not able to identify the specification errors.

If a turning point is hypothesized but its location is unknown, the MI-based procedure proposed by Kwok et al. [6] can be used to fit a linear growth curve model in the data and then identify largest MI for factor loadings of the linear growth factor; however, the procedure requires moderately large sample size (400 and above) and more measurement waves (minimally 8 waves in the study) to have adequate statistical power to detect the turning point. A more powerful alternative to the MI-procedure is the piecewise linear model with unknown turning points estimated based on data. The present findings regarding the performance of the procedure in recovering the mean growth trajectory showed the estimation is highly accurate even if the multivariate normality assumption is violated. Applied researchers are recommended to take advantage of the procedure to correctly identify the turning point in specifying a piecewise linear growth model. Yet, for applied researchers who are interested in the significance test of the turning point and/or the interindividual difference, it is cautioned that the departure from multivariate normality assumption in longitudinal responses tend to inflate the standard error estimates of the turning point and deflate the random effect estimates associated with the growth factors.

Limitations and Future Research Directions

The findings of the study should be considered in light of the limitations and may not be generalized to models and data scenarios that are very different from the ones considered in this study. A limitation to be considered in generalizing the findings of the study is with respect to the level of change in growth rate. As a matter of fact, the change in growth rate can be much larger than what has been considered in our study (e.g., in Kohli and Harring's [5] study, the resulted effect size from the change in growth rate is 5). Another limitation to be considered is that we only examined the impact of a misspecified turning point on growth trait estimation in a priori piecewise linear model, assuming all other parts of the latent growth model were correctly specified. Yet, in real data scenarios, the misspecification of turning point can happen simultaneously with misspecifications in the other parts of the latent growth model (e.g., the misspecification of residual variances across the measurement waves often happens). How the misspecifications in both the turning point and other parts of the latent growth model interact with the design factors considered in present study particularly when the normal distributions were violated is another question that merits research attention.

Finally, the study only considered two-piece linear growth curves connected by one fixed turning point. In reality, developmental trajectories may have a zigzag shape with multiple turning points. In addition, there might be individual differences in the location of the turning points. Hence, further development of the PGCM is needed to model trajectories with multiple unknown turning points and random effects associated with the turning points.

AUTHOR CONTRIBUTIONS

LN and WL jointly conceived and designed the study. LN acquired the data, conducted the analysis and interpretation of the data with the help of WL. LN drafted the manuscript with the conceptual advice from WL. LN and WL worked jointly in revising the manuscript critically for important intellectual content and for the final approval of the version to be published and for the accountability for all aspects of the work to ensure that the questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

ACKNOWLEDGMENTS

The open access publishing fees for this article have been covered in part by the Texas A&M University Open Access to Knowledge Fund (OAKFund), supported by the University Libraries and the Office of the Vice President for Research. The Center for Student Affairs Assessment at University of California, Davis, has also contributed partly to the publishing fees for this article.

REFERENCES

1. Preacher KJ, Wichman AL, MacCallum RC, Briggs NE. *Latent Growth Curve Modeling.* Los Angeles, CA: Sage (2008).
2. Singer JD, Willett JB. *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence.* New York, NY: Oxford University Press (2003).
3. Harring JR, Cudeck R, du Toit SH. Fitting partially nonlinear random coefficient models as SEMs. *Multivariate Behav Res.* (2006) **41**:579–96. doi: 10.1207/s15327906mbr4104_7
4. Khoo S, West SG, Wu W, Kwok O. Longitudinal Methods. In: Eid M, Diener E, editors. *Handbook of Psychological Measurement: A Multimethod Perspective.* Washington, DC: APA (2006). p. 301–17.
5. Kohli N, Harring JR. Modeling growth in latent variables using a piecewise function. *Multivariate Behav. Res.* (2013) **48**:370–97. doi: 10.1080/00273171.2013.778191

6. Kwok O, Luo W, West SG. Using modification indexes to detect turning points in longitudinal data: a monte carlo study. *Struct Equ Model.* (2010) **17**:216–40. doi: 10.1080/10705511003659359
7. Chou C, Yang D, Pentz MA, Hser Y. Piecewise growth curve modeling approach for longitudinal prevention study. *Comput Stat Data Anal.* (2004) **46**:213–25. doi: 10.1016/S0167-9473(03)00149-X
8. Rutter M. Autism research: prospects and priorities. *J Autism Dev Disord.* (1996) **26**:257–75. doi: 10.1146/annurev.psych.52.1.501
9. Hardy SA, Thiels C. Using latent growth curve modeling in clinical treatment research: an example comparing guided self-change and cognitive behavioral therapy treatments for bulimia nervosa. *Int J Clin Health Psychol.* (2009) **9**:51–71. doi: 10.1016/j.eurpsy.2007.01.590
10. Terrera GM, Matthews F, Brayne C. A comparison of parametric models for the investigation of the shape of cognitive change in the older population. *BMC Neurol.* (2008) **8**:16. doi: 10.1186/1471-2377-8-16
11. Dominicus A, Ripatti S, Pedersen N, Palmgren J. *Modelling Variability in Longitudinal Data Using Random Change Point Models.* Research Reports in Mathematical Statistics. Stockholm University (2006). Available online at: http://www.math.su.se/matstat
12. Wang L, McArdle JJ. A simulation study comparison of Bayesian estimation with conventional methods for estimating unknown change points. *Struct Equ Model.* (2008) **15**:52–74. doi: 10.1080/10705510701758265
13. Wood PK, Jackson KM. Escaping the snare of chronological growth and launching a free curve alternative: general deviance as latent growth model. *Dev Psychopathol.* (2013) **25**:739–54. doi: 10.1017/S095457941300014X
14. Preacher KJ, Hancock GR. Meaningful aspects of change as novel random coefficients: a general method for reparameterizing longitudinal models. *Psychol Methods* (2015) **20**:84. doi: 10.1037/met0000028
15. Wu W, Jia F, Kinai R, Little TD. Optimal number and allocation of data collection points for linear spline growth curve modeling: a search for efficient designs. *Int J Behav Dev.* (2019) **41**:550–58. doi: 10.1177/0165025416644076
16. Compton DL, Fuchs LS, Fuchs D, Lambert W, Hamlett C. The cognitive and academic profiles of reading and mathematics learning disabilities. *J Learn Disabil.* (2012) **45**:79–95. doi: 10.1177/0022219410393012
17. Steiger JH. Structural model evaluation and modification: an interval estimation approach. *Multivariate Behav. Res.* (1990) **25**:173–80. doi: 10.1207/s15327906mbr2502_4
18. Jöreskog KG, Sörbom D. *LISREL 7: A Guide to the Program and Applications*. Chicago, IL: SPSS (1988).
19. Bentler PM. Comparative fit indexes in structural models. *Psychol. Bull.* (1990) **107**:238–46. doi: 10.1037/0033-2909.107.2.238
20. Tucker LR, Lewis C. A reliability coefficient for maximum likelihood factor analysis. *Psychometrika* (1973) **38**:1–10. doi: 10.1007/BF02291170
21. Hu L, Bentler PM. Fit indices in covariance structure analysis: sensitivity to underparameterized model misspecification. *Psychol Methods* (1998) **3**:424–53. doi: 10.1037/1082-989X.3.4.424
22. Hu L, Bentler PM. Cutoff criterion for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struct Equ Model.* (1999) **6**:1–55. doi: 10.1080/10705519909540118
23. Fan X, Sivo SA. Sensitivity of fit indexes to mis-specified structural or measurement model components: rationale of two-index strategy revisited. *Struct Equ Model* (2005) **12**:343–67. doi: 10.1207/s15328007sem1203_1
24. Fan X, Sivo SA. Sensitivity of fit indices to model misspecification and model types. *Multivariate Behav. Res.* (2007) **42**:509–29. doi: 10.1080/00273170701382864
25. Chen F, Curran PJ, Bollen KA, Kirby J, Paxton P. An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. *Sociol. Methods Res.* (2008) **36**:462–94. doi: 10.1177/0049124108314720
26. Wu W, West SG, Taylor AB. Evaluating model fit for growth curve models: integration of fit indices from SEM and MLM frameworks. *Psychol. Methods* (2009) 14:183. doi: 10.1037/a0015858
27. Wu W, West SG. Sensitivity of fit indices to misspecification in growth curve models. *Multivariate Behav. Res.* (2010) **45**:420–52. doi: 10.1080/00273171.2010.483378
28. Leite WL, Stapleton LM. Detecting growth shape misspecifications in latent growth models: an evaluation of fit indexes. *J Exp Educ.* (2011) **79**:361–81. doi: 10.1080/00220973.2010.509369
29. Widaman KF, Thompson JS. On specifying the null model for incremental fit indices in structural equation modeling. *Psychol Methods* (2003) **8**:16–37. doi: 10.1037/1082-989X.8.1.16
30. Hutchinson SR. Univariate and multivariate specification search indices in covariance structure modeling. *J Exp Educ.* (1993) **61**:171–81. doi: 10.1080/00220973.1993.9943859
31. Saris WE, Satorra A, Sörbom D. The detection and correction of specification errors in structural equation models. In: Clogg CC, editor, *Sociological Methodology*, San Francisco, CA: Jossey-Bass (1987). p. 105–129.
32. Kaplan D. The impact of specification error on the estimation, testing, improvement of structural equation models. *Multivariate Behav Res.* (1988) **23**:69–86. doi: 10.1207/s15327906mbr2301_4
33. Chou CP, Bentler PM. Invariant standardized estimated parameter change for model modification in covariance structure analysis. *Multivariate Behav. Res.* (1993) **28**:97–110. doi: 10.1207/s15327906mbr2801_6
34. Whittaker TA. Using the modification index and standardized expected parameter change for model modification. *J Exp Educ.* (2012) **80**:26–44. doi: 10.1080/00220973.2010.531299
35. Saris WE, Satorra A, Van der Veld WM. Testing structural equation models or detection of misspecifications?. *Struct Equ Model.* (2009) **16**:561–82. doi: 10.1080/10705510903203433
36. Chou CP, Bentler PM. Model modification in covariance structure modeling: a comparison among likelihood ratio, lagrange multiplier, Wald tests. *Multivariate Behav Res.* (1990) **25**:115–36. doi: 10.1207/s15327906mbr2501_13
37. Luijben TC, Boomsma A. Statistical guidance for model modification in covariance structure analysis. *Compstat* (1988) **1988**:335–40. doi: 10.1007/978-3-642-46900-8_46
38. MacCallum RC. Specification searches in covariance structure modeling. *Psychol. Bull.* (1986) **100**:107–20. doi: 10.1037/0033-2909.100.1.107
39. Raudenbush SW, Liu XF. Effects of study duration, frequency of observation, and sample size on power in studies of group differences in polynomial change. *Psychol. Methods* (2001) **6**:387. doi: 10.1037/1082-989X.6.4.387
40. Vale CD, Maurelli VA. Simulating multivariate nonnormal distributions. *Psychometrika* (1983) **48**:465–71. doi: 10.1007/BF02293687
41. Micceri T. The unicorn, the normal curve, and other improbable creatures. *Psychol Bull.* (1989) **105**:156. doi: 10.1037/0033-2909.105.1.156
42. Bauer DJ, Curran PJ. Distributional assumptions of growth mixture models: implications for overextraction of latent trajectory classes. *Psychol Methods* (2003) **8**:338. doi: 10.1037/1082-989X.8.3.338
43. SAS Institute (2016). *SAS, Release 9.4 [Computer software]*. Cary, NC: SAS Institute.
44. Muthén LK, Muthén BO. *M plus User's Guide Seventh Edition*. (1998–2016). Los Angeles, CA: Muthén & Muthén.
45. Boomsma A. *On the Robustness of LISREL (Maximum Likelihood Estimation) Against Small Sample Size and Non-Normality*. Doctoral dissertation, University of Groningen (1983).
46. McIntosh CN. Rethinking fit assessment in structural equation modeling: a commentary and elaboration on Barrett (2007). *Pers Individ Diff.* (2007) **42**:859–67. doi: 10.1016/j.paid.2006.09.020
47. Yuan K. Fit indices versus test statistics. *Multivariate Behav Res.* (2005) **40**:115–48. doi: 10.1207/s15327906mbr4001_5
48. Nevitt J, Hancock GR. Improving the root mean square error of approximation for nonnormal conditions in structural equation modeling. *J Exp Educ.* (2000). **68**:251–68. doi: 10.1080/00220970009600095

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer KJK and handling Editor declared their shared affiliation.

APPENDIX

Mplus Code for Fitting Newly Proposed PGCM with an Unknown Turning Point

```
Title: Newly Proposed PGCM;

Data:
    File is “C:\tpoint_4\data.txt”;

Variable:
    Names are ID t1-t7;
    Usevariables are t1-t7;

Analysis: Estimator=MLR;

MODEL:
    w1 BY t1-t7@1;
    w2 BY t1@0 t2@1 t3@2 t4@3 t5@4 t6@5 t7@6;
    w3 BY t1*0 (p1);
    w3 BY t2-t7 (p2-p7);

    w1; w2; w3;
    w1 WITH w2*0;
    w1 WITH w3*0;
    w2 WITH w3*0;

    [t1-t7@0];
    t1-t7*1;

    [w1](mw11);
    [w2](mw21);
    [w3](mw31);

MODEL CONSTRAINT:

      NEW(gam1*2.5 b11*2.0 b21*0.5 b41*0.3); ! The starting values set to be around the true values;
    p1 = (sqrt((0-gam1)^2));
    p2 = (sqrt((1-gam1)^2));
    p3 = (sqrt((2-gam1)^2));
    p4 = (sqrt((3-gam1)^2));
    p5 = (sqrt((4-gam1)^2));
    p6 = (sqrt((5-gam1)^2));
    p7 = (sqrt((6-gam1)^2));
    b11 = mw11+mw31*gam1;
    b21 = mw21-mw31;
    b41 = mw21+mw31;

OUTPUT:
```

frontiers in Psychology

ORIGINAL RESEARCH
published: 27 March 2018
doi: 10.3389/fpsyg.2018.00349

The Optimal Starting Model to Search for the Accurate Growth Trajectory in Latent Growth Models

Minjung Kim[1]*, *Hsien-Yuan Hsu*[2], *Oi-man Kwok*[3] *and Sunmi Seo*[4]

[1] Quantitative Research, Evaluation, and Measurement, Department of Educational Studies, The Ohio State University, Columbus, OH, United States, [2] Children's Learning Institute, University of Texas Health Science Center at Houston, Houston, TX, United States, [3] Department of Educational Psychology, Texas A&M University, College Station, TX, United States, [4] Department of Psychology, University of Alabama, Tuscaloosa, AL, United States

This simulation study aims to propose an optimal starting model to search for the accurate growth trajectory in Latent Growth Models (LGM). We examine the performance of four different starting models in terms of the complexity of the mean and within-subject variance-covariance (V-CV) structures when there are time-invariant covariates embedded in the population models. Results showed that the model search starting with the fully saturated model (i.e., the most complex mean and within-subject V-CV model) recovers best for the true growth trajectory in simulations. Specifically, the fully saturated starting model with using ΔBIC and ΔAIC performed best (over 95%) and recommended for researchers. An illustration of the proposed method is given using the empirical secondary dataset. Implications of the findings and limitations are discussed.

Keywords: latent growth models, latent curve models, growth analysis, model selection, specification search, starting model, model building, growth curve

OPEN ACCESS

Edited by:
Prathiba Natesan,
University of North Texas,
United States

Reviewed by:
Akihito Kamata,
Southern Methodist University,
United States
Ke-Hai Yuan,
University of Notre Dame,
United States

****Correspondence:***
Minjung Kim
mjkim.epsy@gmail.com

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *16 August 2017*
Accepted: *02 March 2018*
Published: *27 March 2018*

Citation:
Kim M, Hsu H-Y, Kwok O and Seo S (2018) The Optimal Starting Model to Search for the Accurate Growth Trajectory in Latent Growth Models. Front. Psychol. 9:349. doi: 10.3389/fpsyg.2018.00349

INTRODUCTION

Longitudinal data has been widely used in many research areas including medical, education, and psychology. One of the major questions when using longitudinal data is often on the change of the measured variables over time, such as: *are parental control and knowledge for their children declining over time?* (Keijsers and Poulin, 2013); *what are the developmental trajectories for adolescents' empathic concern associated with pubertal status?* (Van der Graaff et al., 2014). Most educational and psychological researchers are interested in not only the accurate growth, but also the factors/covariates (e.g., gender, involvement in peer-oriented leisure activities) accounting for the variation of growth trajectory among participants (Crockett and Beal, 2012; Titzmann et al., 2014). Latent growth models (LGM; also called latent growth curve models) have been increasingly popular in longitudinal studies given that the latent growth models allow researchers to take into account the between-individual differences as well as within-individual differences over time (Meredith and Tisak, 1990; Preacher et al., 2008; Duncan et al., 2013).

In longitudinal data analysis under LGM, many studies have devoted to optimally model the overall shape of the growth trajectories for all subjects based on the hypothesized model (Duncan et al., 1994; Hancock and Lawrence, 2006; Blozis, 2007). When there is no hypothesized theory, however, researchers may use exploratory approach to search for the optimal growth shape based on their data. Visual inspection using graphical function in statistical software (e.g., empirical growth plot) can be one viable approach to start with, but it is more suitable with a subset of sample rather than with a large sample data (Singer and Willett, 2003). While the traditional model building approach has been employed for decades, under this circumstance, there have been extensive efforts to suggest the model specification search strategy for the optimal shape of growth trajectory (Leite and Stapleton, 2011; Liu et al., 2012; Kim et al., 2016; Whittaker and Khojasteh, 2017). Under the framework of LGM, model specification search can be conducted in terms of the mean structure (i.e., shape of the overall changing pattern) and variance-covariance (V-CV) structure consisting of growth factor V-CV (i.e., variations across the individual growth trajectories) and residual V-CV structure (i.e., variations within the individual growth trajectories). Previous research has consistently found that the saturated residual variance-covariance structure (i.e., freely estimating the variance and covariances of repeated measures) has promising performance when searching for the accurate growth shape in simulations (Wu and West, 2010; Kim et al., 2016). However, existing recommendation has been made upon previous simulations, assuming that all growth latent factors are exogenous variables in the population models. That is, no studies have investigated whether existing recommendation is still applicable to the case that growth latent factors are both exogenous and endogenous variables at the same time. When the possible covariates are excluded in the model, the latent growth models is regarded to be *misspecified* given that the paths from the covariates to the growth factors are constrained to be 0. When the influential covariates are existing but not considered in the step of searching for the accurate growth trajectory, little is known about (a) which starting model performs best in searching for the optimal growth trajectory and (b) which model selection criteria can be used to successfully search for the best growth trajectory.

In the present study, we aim to investigate the optimal model search strategy for finding for the accurate growth shape in simulations when there is a significant covariate associated with the growth trajectory. Specifically, we focus on time-invariant covariates (e.g., gender, years of education, ethnicity) in the current study. Under the framework of LGM, we employ the four different starting models in terms of the complexity of mean and residual variance structure following the previous research by Kim et al. (2016): (1) the simplest mean and the error variance structure, (2) the most complex mean and the simplest residual variance structure, (3) the simplest mean and the most complex error variance structure, and (4) the most complex mean and the error structure. Specifically, we examine (1) which starting model performs best in model specification search, (2) which model evaluation index shows successful performance in finding the population growth shape, and propose the optimal model search strategy given the results of the two research questions. We use a Monte Carlo simulation study to investigate the effectiveness of different starting models on the search for the correct mean trajectory. An illustrative example is also presented to apply the model search strategy.

Mean and Residual Variance (Variance-Covariance) Structures in LGM

There are three model components in LGM: mean structure, between-subject variance-covariance (V-CV) structure, and within-subject V-CV structure. A general model formulation in LGM can be written as:

$$\mathbf{y} = \tau_y + \Lambda_y \eta + \varepsilon, \tag{1}$$

where y refers to a vector of outcome variables (t × 1, where t is the number of repeated measures), τ refers to a vector of intercepts of ys (t × 1; typically fixed to zero for model identification purpose), Λ represents a factor loading matrix for ys (t × p, where p is the number of latent growth factors), η is a vector of latent growth factors (p × 1), and ε represents a vector of errors for each y across the repeated measures (t × 1). η can be further written as follows:

$$\eta = \alpha + \Gamma_\eta w + \zeta, \tag{2}$$

where α contains the vector of population initial status and growth parameters (e.g., $\begin{bmatrix}\alpha_0\\ \alpha_1\end{bmatrix}$ for a linear growth model), Γ represents a matrix of regression coefficients of time-invariant covariate w, and ζ represents the deviation of the corresponding individual values from the mean estimates of those growth factors, respectively. Mean structure is the expected value of y [i.e., $E(y) = \Lambda_y\alpha + \Lambda_y\Gamma_\eta E(w)^1$], which represents the average growth trajectory and the covariate effect on the change rates. In the current study, we aim to correctly search for the structure of growth shape (i.e., $\Lambda_y\alpha$) while omitting the covariate effect part [i.e., $\Lambda_y\Gamma_\eta E(w)$] by using an unconditional model in the search procedure. The variance-covariance (V-CV) of y in Equation (1) can be written as:

$$V(y) = \Sigma = \Lambda_y \Psi \Lambda_y' + \Theta_\varepsilon, \tag{3}$$

where Ψ is a p×p matrix containing the variance and covariances of the growth related latent factors; Λ'_y is the transpose of the Λ matrix which captures the overall pattern of change, and Θ_ε represents the matrix of variances and covariance among the errors (or unique factors). In other words, between-subject V-CV is captured by Ψ matrix, representing the differences on the intercepts and growth shapes among the subjects. In the current study, we focus on the specification of within-subject V-CV structure given that more complex structure and assumptions are associated with the within-subject V-CV components compared to between-subject V-CV structure (Kim et al., 2016). Within-subject V-CV structure (also called residual variance structure throughout this paper) is the variance and covariances of the repeated measures for each individual (t × t matrix of Θ_ε), which captures the deviations of the observed variables from a vector of expected ys.

Model Building Process in Methodological Studies

There have been a number of debates on how to search for the optimal growth shape in longitudinal data analysis. When there is no hypothesized growth shape in the absence of theory,

[1] A vector of τ_y is fixed to zero for model identification purpose.

there are two commonly used starting points in terms of the mean structure: the simplest mean structure (i.e., intercept-only model) and the most complex mean structure (i.e., highest possible polynomial growth model). The simplest intercept-only model has been frequently used for model search process in the longitudinal data analysis given that they follow the classic model search strategy provided by Raudenbush and Bryk (2002) from their classical book of hierarchical linear modeling.

Likewise, under the framework of multilevel model, Singer and Willett (2003) suggested starting from the unconditional model where there is a time-associated factor (e.g., age, year, months) but no other factors or covariates in the model. Similarly, McCoach and Kaniskan (2010) have demonstrated a model building method by starting with an unconditional linear growth model followed by adding time-varying covariates. In their study, the empirical data from 277 elementary school students over four time points are used for the demonstration.

Ryoo (2011) has conducted a simulation study for a model building approach and recommended of using the simplest mean structure model with no covariates as the starting point (i.e., intercept-only model) to search for the true growth shape. In his simulation study, six covariates are included for data generation while those are not included at the first step of model selection process. Static predictors of growth trajectories are introduced into the model at the last step after selecting the proper growth (mean) structure. Results show that the step-up (i.e., starting from the simplest mean structure) approach performs well to search for the true growth shape. Meanwhile, the error variance structure has not been discussed in the study and the default structure (i.e., simplest Identity structure) has been used for all simulation conditions.

On the other hand, under the framework of LGM, Mayer et al. (2012) has illustrated a 3-step model building process using a quadratic growth model as an example to show how to define the latent growth components in longitudinal data analysis. According to Mayer et al. (2012), based on a measurement model formulated at Step 1, specifying the saturated (most complex) mean structure is recommended for a starting model at Step 2 while covariates predicting growth components are added at Step 3.

In most studies, however, model specification for the variance-covariance structure part has been often disregarded in the model building process because it seldom impacts the shape of the growth trajectory itself (Kwok et al., 2007). However, the impact of ignoring the error variance structure gets more severe when conducting a model search because it may end up selecting an inaccurate growth shape as the best fitting model. Recently published study by Kim et al. (2016) shows that specifying the simplest within-subject V-CV structure, which is the default error structure in many statistical software, is less likely to select the optimal growth shape as the best fitting model. In their simulations, the average recovery rate for finding the population growth shape is $<50\%$ when using the simplest error variance structure, while it is above 85% when saturating the residual V-CV structure with using certain model evaluation criteria (e.g., LRT, ΔAIC, and ΔBIC). In their study, only unconditional models without covariates are used as a population model. There are no studies, at our knowledge, investigating the model specification search for the population growth shape in LGM, considering both mean and error variance structure when covariates are regressed on the growth factors.

Applied Studies Using LGM

Many applied studies employing latent growth models under the multilevel modeling framework typically use the simplest within-subject V-CV structure (i.e., Identity [ID]; constant variance across repeated measures without allowing any covariance between the measures) because it is the default error variance structure in MLM software (e.g., SPSS MIXED, SAS PROC MIXED, HLM). Although there are published tutorials available for how to change the default within-subject error variance structure (or level-1 residual structure in MLM framework) (Quené and Van den Bergh, 2004), modifying the residual structure has been rarely considered in most applied research.

We reviewed substantive studies published in *Developmental Psychology* between 2010 and 2016 and found 37 studies[2] employing the latent growth (or growth curve) models for the longitudinal data analysis. Among 37 studies, 15 studies specified a linear growth model with no search procedure due to the limited number of repeated measures (i.e., 3 waves). Among 22 of 37 studies containing 4 or more waves of data, 14 studies (63.6%) conducted a model comparison to find the best fitting growth trajectory while 8 studies directly specified their hypothesized growth shape (i.e., linear growth model for 7 studies and piece-wise growth model for one study). Among 14 studies conducting a model comparison, 8 studies contained 4 waves of data and they compared a linear growth model to a non-linear growth model (e.g., quadratic growth model). Among the rest of 6 studies, which conducted a model specification search with more than 4 waves of data, three studies reported the fit statistics (e.g., chi-square difference test, CFI, RMSEA, and SRMR) for all compared models. Nevertheless, none of studies reported the information regarding the specification of residual variance structure during the model search procedure. For the selected final model, majority of the studies (86.5%) directly specified the simplest residual variance structure without considering other types of V-CV structures. As shown in the reviewed literatures, there is a lack of consensus for using a model building approach in latent growth models to search for the optimal growth trajectory.

STUDY AIMS

Our goal is to propose a universal starting model to search for the best-representing growth shape for the data regardless of the *true* population mean structure because, in reality, we do not know the true or accurate growth trajectory. We followed Kim et al. (2016) to set up the four starting models in terms of the mean and the residual structures in LGM. **Figure 1** presents the four possible starting models for 4 wave data as an example: (1) the simplest mean (intercept-only) with the simplest

[2] A list of studies is provided on the first author's website as an Appendix.

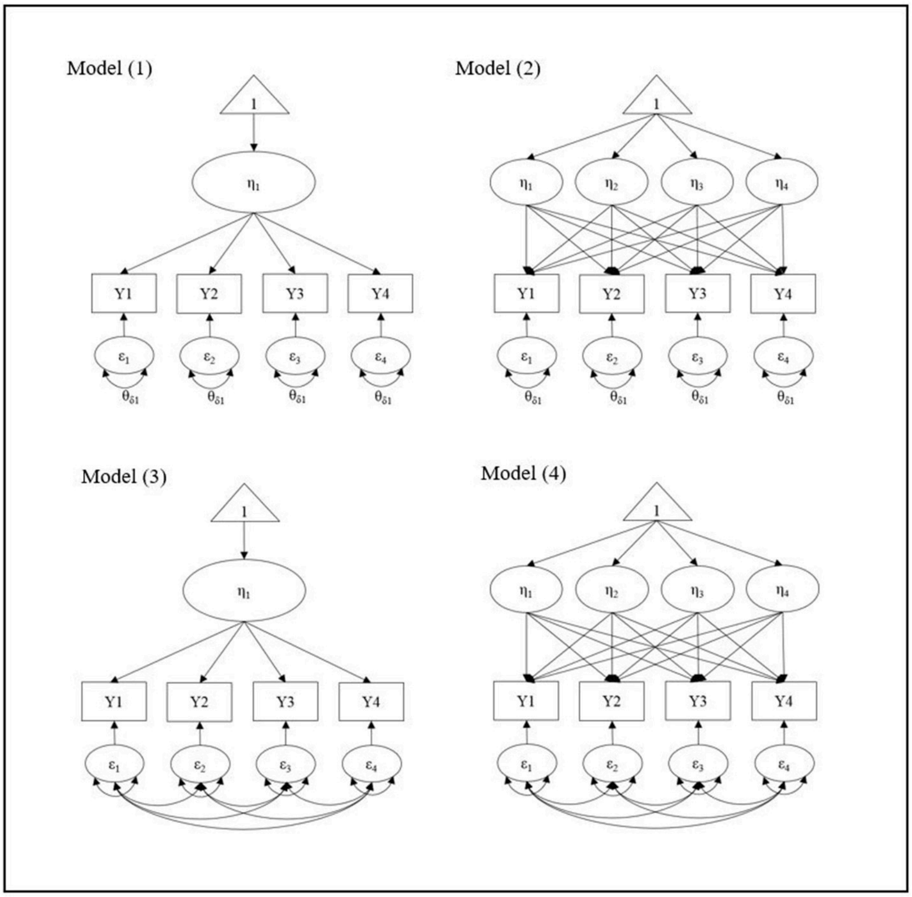


FIGURE 1 | Four starting models for 4 wave data.

ID error variance structure, (2) the most complex mean (e.g., highest possible polynomial growth term) and the simplest ID structure, (3) the simplest mean and the most complex UN error variance structure, and (4) the most complex mean and the error structure. We extend the previous study to consider more general conditions, in which there is a covariate effect on the growth trajectories. While it has been found that saturating the within-subject V-CV structure performs successfully to search for the true growth trajectory without considering covariates (Wu and West, 2010; Kim et al., 2016), the starting point for the mean structure has shown no consistent results. Given that the previous research used no covariates for the true model setting, we expand it to more general model with covariates and examine whether the consistent results can be found in more general conditions. We have two specific research questions in the current study.

> *Q1: Which starting model performs best in searching for the correct growth trajectory?*

We examine the performance of four unconditional growth starting models to search for a population growth shape under the LGM framework. Based on the previous research, we hypothesize that the model specified with the most complex residual variance structure will perform successfully in searching for the growth shape. Given that the starting point for the mean structure has shown inconsistent results, we specifically interested in: *Does specification in mean structure (the most complex vs. the simplest) affect the recovery rate for detecting the true growth trajectory?*

> *Q2: Which model selection criteria performs successfully to search for the true growth trajectory?*

We use six commonly used model evaluation criteria (i.e., LRT, ΔCFI, ΔRMSEA, ΔSRMR, ΔAIC, and ΔBIC) with two different model building approaches (i.e., step-up and top-down). We expect that LRT and two information criteria (i.e., ΔAIC and ΔBIC) will outperform the other fit indices based on the previous research finding (Kim et al., 2016).

METHODS: SIMULATION STUDY

Data Generation

Data are generated using Mplus7.1 (Muthén and Muthén, 1998-2012) with a multivariate normal distribution. We have four major design factors in this simulation study: (a) 2 number of waves and mean structure (4 and 8 for linear and quadratic model, respectively), (b) true residual variance structure [ID, UN(1), and AR(1)][3], (c) 3 covariate effect sizes (0.1, 0.3, and 0.5), and (d) 3 sample sizes (100, 210, and 390), yielding a total of 54,000 datasets (2 × 3 × 3 × 3 × 1,000 replications). A thousand replications per simulation condition is reasonable for a simulation study in SEM given that many previous research have used equal to or fewer than 1,000 replications. **Figure 2** shows an example of population model with 4 waves of data, which is a linear growth model with the UN(1) error variance structure. More details in simulation conditions for each design factor are described in the following section.

Number of Waves and Mean Structure

We have used two conditions for the number of repeated measures, 4 and 8, for building up the population model of the growth trajectory. The number is based on reviewing the substantive studies published in developmental psychology between 2010 and 2016 as well as the previous simulation study (Kwok et al., 2007; Kim et al., 2016). The average number of waves used in longitudinal data analysis is 4.4 with a standard deviation of 1.6. Among a total of 37 reviewed studies employing the latent growth models, 24 studies have modeled a linear trend to analyze their data while 9 studies have used a quadratic growth trajectory to best represent their data. The rest of 4 studies have modeled their data other than linear and quadratic (e.g., piecewise growth model, factor loading freed non-linear model). Therefore, we have set up a population model of 4 waves of data to be a linear growth and 8 waves of data (i.e., approximately 2 standard deviations above the mean number of waves) to be a quadratic growth. Population values for both growth trajectories are set up to be a medium effect, which have been employed in the previous simulation studies (Kwok et al., 2007; Kim et al., 2016).

Residual Variance Structure

For generating the datasets representing the population model, we have used three types of variance-covariance structures, Identity (ID), Autoregressive [AR(1)], and banded main diagonal [UN(1)][4], which are commonly used in many longitudinal studies as well as simulation studies (Kwok et al., 2007; Kim et al., 2016). Among the 37 reviewed studies, 32 studies (86.5%) have used the simplest error V-CV structure (i.e., ID). Two studies have allowed a correlation between the error terms and 3 studies have estimated the time-specific variances [UN(1)]. The residual variances of the measurement waves (i.e., θ_δ) were all set to be 1.00 for both ID and AR(1) structures, which was a common practice in power analysis and simulation studies. Following the prior simulation studies on residual variance structure, the autocorrelation coefficient, ρ, was set to be 0.50 for AR(1) structure. For the UN(1) structure, all the covariances were set to zero while the residual variance of the first time point was set to 1.00 and the following residual variances were set to be the power function of $\rho = 0.80$ (i.e., $\sigma_1^2 = 1.00$, $\sigma_2^2 = 0.80$, $\sigma_3^2 = 0.64$, and $\sigma_4^2 = 0.51$ for 4 waves of data; $\sigma_1^2 = 1.00$, $\sigma_2^2 = 0.80$, $\sigma_3^2 = 0.64$, $\sigma_4^2 = 0.51$, $\sigma_5^2 = 0.41$, $\sigma_6^2 = 0.33$, $\sigma_7^2 = 0.26$, and $\sigma_8^2 = 0.21$ for 8 waves of data), assuming that the reliability of the measurement increases over time (Grimm and Widaman, 2010).

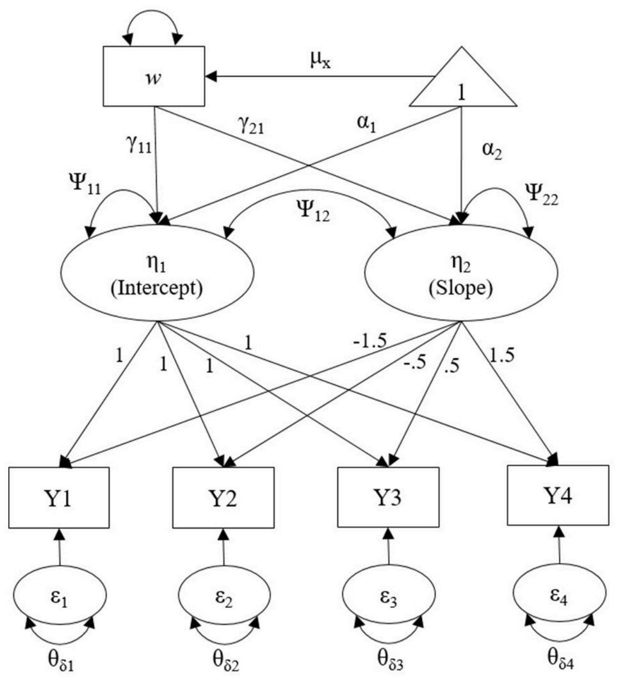


FIGURE 2 | Linear growth model with UN(1) error variance structure with 4 waves of data to generate the true population model with a single covariate.

Between-Subject V-CV Structure

We adopted the population parameters for the between-subject V-CV structure from the previous simulation studies in LGM (Kwok et al., 2007; Kim et al., 2016) Given that intercept variance has generally been larger than the variation of the change in growth in longitudinal studies (Raudenbush and Xiao-Feng, 2001), the total variance of Ψ_{11} was set to 0.20 while both Ψ_{22} and Ψ_{33} were set to 0.10 constantly for all conditions. The elements in the matrix were set to:

$$\Psi_{\text{Linear}} = \begin{bmatrix} \Psi_{11} & \Psi_{12} \\ \Psi_{21} & \Psi_{22} \end{bmatrix} = \begin{bmatrix} 0.20 & 0.05 \\ 0.05 & 0.10 \end{bmatrix}$$

and

$$\Psi_{\text{Quadratic}} = \begin{bmatrix} \Psi_{11} & \Psi_{12} & \Psi_{13} \\ \Psi_{21} & \Psi_{22} & \Psi_{23} \\ \Psi_{31} & \Psi_{32} & \Psi_{33} \end{bmatrix} = \begin{bmatrix} 0.20 & 0.05 & 0.05 \\ 0.05 & 0.10 & 0.035 \\ 0.05 & 0.035 & 0.10 \end{bmatrix},$$

with the correlations (i.e., $r = \frac{\Psi_{xy}}{\sqrt{\Psi_{xx}\Psi_{yy}}}$, $x \neq y$) setting as 0.35 for all the pairs of the elements in the Ψ matrix (Kwok et al., 2007). Based on the covariate effect sizes, the size of the variance

[3] ID=Identity; UN(1)=Banded main diagonal; AR(1)=Autoregressive.

[4] $\text{ID} = \sigma^2 \begin{bmatrix} 1 & & & \\ 0 & 1 & & \\ 0 & 0 & 1 & \\ 0 & 0 & 0 & 1 \end{bmatrix}$; $\text{AR}(1) = \sigma^2 \begin{bmatrix} 1 & & & \\ \rho & 1 & & \\ \rho^2 & \rho & 1 & \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$; $\text{UN}(1) = \begin{bmatrix} \sigma_1^2 & & & \\ 0 & \sigma_2^2 & & \\ 0 & 0 & \sigma_3^2 & \\ 0 & 0 & 0 & \sigma_4^2 \end{bmatrix}$.

and covariance of the growth associated factors was adjusted to consider the explained variance given the covariate.

Covariate Effect Size

To ease the understanding of mechanism under the model formulation with covariates, we have used a simple model with a single covariate in this simulation. Because adding more predictors is a function of increasing the total effect size, which decreases the size of residual (unexplained) variance, we have used three different sizes of covariate effects on the mean growth structure: 0.1 (small), 0.3 (medium), and 0.5 (large) while keeping a single covariate. The covariate, *w*, is generated to have a mean of 0 and a standard deviation of 1 with a normally distributed variance. The covariate effects are equally regressed on each growth-related term. For example, for the true linear growth model condition, a covariate has the same coefficient on the intercept and the slope.

Sample Size

We have used three sample sizes, which are 100, 210, and 390, for small, medium, and large sample size conditions, respectively, following the previous simulation study (Kim et al., 2016). Thus, the total number of observations ranged from 400 (4 wave × 100 subjects) to 3,120 (8 wave × 390 subjects). We expect that the model stability will increase as sample size increases, indicating better chance of finding the correct mean trajectory.

Evaluation Criteria

We have evaluated three types of model selection criteria on the performance of finding the correct mean growth trajectory: (a) Likelihood Ratio Test (LRT), (b) ΔGoodness of Fit Indices [ΔGFI; i.e., Comparative Fit Index (CFI), Root Mean Square Error of Approximation (RMSEA), and Standardized Root Mean Residual (SRMR)], and (b) ΔInformation Criteria [ΔIC; i.e., Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC)]. These model selection criteria are commonly used by applied researchers because most statistical software for SEM provide these fit statistics in the output (e.g., Mplus, Lisrel, Amos). The difference in GFI and IC between the two competing models (i.e., constrained model vs. relaxed model) is calculated for each of the model fit index. Previous research has shown promising results for using both LRT and ΔIC in model specification search (Kim et al., 2016), whereas the ΔGFI showed inconsistent results. Following Kim et al. (2016), we have used the six model fit indices for evaluation criteria.

For information criteria (i.e., AIC and BIC), we have set up an absolute value to select a better fitting model over the competing model. Burnham and Anderson (2004) suggested using 4 for AIC to decide whether a model fit is significantly improved by adding additional parameter. Similary, Raftery (1995) suggested using 2 for BIC to compare the models. In other words, when the difference on the information criteria between the two competing models is minimal (i.e., ΔAIC < 4; ΔBIC < 2), we have selected the simpler model even when the more complex model showed smaller value of AIC and BIC. Similarly, we have adopted more stringent cutoff criteria for the GFIs proposed by Chen (2007). When the difference on CFI between the simpler and more complex model is <0.01, the simpler model was selected over the more complex model. The cutoff for ΔRMSEA and ΔSRMR are 0.015 and 0.01, respectively.

Model Search Process

For each dataset, four sets of model search procedure have been conducted using the four different starting models. Step-up refers to starting from the simplest mean structure (i.e., intercept-only model) by adding one more growth related factors at a time. For example, the intercept-only model with ID structure is compared to the linear growth model with the same residual variance structure using each of six different model evaluation criteria. If the model is significantly improved by adding a linear growth term, then the linear growth model is compared to the quadratic growth model in the next step, and so on. When the model is not improved any more, model search process is stopped and the simpler model between the two competing models is selected to be the optimal growth trajectory. If the selected growth trajectory is matched with the true (generated) growth structure, "hit" is coded as 1, while the incorrect growth trajectory is coded as 0. Since this process is independently conducted by six model evaluation criteria, the hit rates are varied across the model evaluation criteria. In a similar manner, top-down refers to starting from the most complex mean structure (i.e., cubic growth model for 4 wave; sextic (6th-order polynomial) growth model for 8 wave) by removing the highest growth related factor at a time. If the more complex model significantly fits better to the data, search has been stopped and the more complex model has been selected as the best fitting model.

Dependent Variable

The primary dependent variable was the hit rate of the true mean model being successfully identified by the model selection indices across the different starting models. For this dependent variable, correct model recovery was coded as a binary variable (i.e., 0 for a miss and 1 for a hit) for all replicates by all conditions. The hit rate (i.e., percentage of replicates reaching the true mean model) was summarized according to the performance of different starting models and model selection fit indices.

RESULTS: SIMULATION STUDY

Before using the model search process, we first analyzed the correctly specified model in terms of both mean and within-subject V-CV structures to validate the data generation process. Results show that all simulations for linear and quadratic growth models with the corresponding true error variance structures [i.e., ID, AR(1), and UN(1)] are properly converged with the accurate parameter estimates indicating that the data were adequately generated. Next, for each true model, four different starting models have been utilized to search for the true mean structure: (1) the simplest mean (intercept-only) with the simplest ID error variance structure, (2) the most complex mean (e.g., highest possible polynomial growth term) and the simplest ID structure, (3) the simplest mean and the most complex UN error variance structure, and (4) the most complex mean and the error variance structure. We present the results of our simulation studies by two research questions.

Which Starting Model Performs Best in Searching for the Correct Growth Trajectory?

Table 1 presents the average hit rates (i.e., percentage of replicates reaching the true growth shape) across all six model fit evaluation criteria when using the four different starting models. Although each fit index is used independently for model search, we average the hit rates of all six fit indices to clearly compare the performance of the four starting models corresponding to the research question. The first three columns provide the information regarding the analyzed model using different starting points[5]. For the within-subject V-CV structure, ID [identity] is the simplest structure while UN [unstructured] is the most complex structure. For the mean structure in the next column, step-up refers to starting with the simplest mean (i.e., intercept-only model) while top-down refers to starting with the most complex mean model (i.e., cubic growth model for 4 wave data; 6th-order polynomial growth model[6] for 8 wave data). The next two columns give the information about the true model conditions for the covariate effect size and sample size. The 6th to 11th columns report the average hit rates under six different true model conditions.

As shown in **Table 1**, starting model (4), which is specified with the most complex mean and residual variance structure, performs best in searching for the population growth trajectory. The average hit rate is 82.3% across all simulation conditions and all model selection criteria. The average hit rates for six different population models range between 75.6 and 87.3% indicating relatively stable performance across all simulation conditions. As covariate effect size and sample size increase, the percentage of finding the true growth shape slightly increases when using the starting model (4). Following Model (4), Model (3) that uses the intercept-only with the most complex UN structure as the starting point shows 70.5% of average hit rate with a range between 47.8 and 90.8%. While Model (3) performs relatively well for searching for the linear growth model (ranged between 88.3 and 90.8%), hit rates are substantially decreased for the quadratic growth model (ranged between 47.8 and 53.9%). As shown in **Table 1**, as covariate effect size increases, hit rates for Model (3) decreases across different sample size conditions and error variance structures. Results for each fit index show that the fit statistic difference between the intercept-only model and linear growth model is minimal, which leads to select the intercept-only model as the better fitting model than the linear growth model. Although Model (3) outperforms Model (4) for the true linear growth model, it shows unstable results for the true quadratic mean structure, which indicates that Model (3) is sensitive to the true mean structure while Model (4) is relatively robust to the true growth shape. More specifically, hit rates for Model (3) substantially decreases when the sample size becomes smaller and covariate effect size gets larger.

Meanwhile, Model (1), which is the most commonly used starting model in practice, shows the worst performance in searching for the accurate growth shape with overall average hit rate of 50.1% (ranged between 21.0 and 81.5%). Only when the true model is a linear growth model with the ID structure, Model (1) shows a good performance (81.5% hit rate). Given that not only the true mean structure is adjacent to the starting mean structure (i.e., a linear growth model and an intercept-only model) but also the V-CV structure is correctly specified (i.e., ID), it can be well expected that Model (1) performs successfully under this specific condition. Similarly, Model (2) (i.e., the most complex mean with the simplest V-CV structure model) shows no promising results in searching for the true mean structure with the average hit rate of 53.1% (ranged between 22.9 and 90.9%). Model (2) shows a good performance only when the true V-C structure is the true ID structure; the average hit rates are 88.0 and 90.9% for the linear and quadratic growth model, respectively. However, when the true V-CV structure is not ID but UN(1) or AR(1), both Model (1) and (2) show poor performance in detecting the true shape of the growth. Notably, Model (1) and (2) perform worse as sample size increases, which can be an evidence of the unstable model results (Kim et al., 2016).

Which Model Selection Index Performs Well in Searching for the Accurate Growth Shape?

Table 2 presents the average hit rates for six model selection fit indices across all simulation conditions. As shown in the table, ΔBIC shows the highest average hit rate (84.4%) across all simulation conditions followed by ΔAIC (average hit rate of 73.3%). The average hit rate of LRT across all simulations is 67.5% followed by ΔSRMR (57.4%), ΔCFI (53.6%), and ΔRMSEA (48.2%). Specifically, ΔBIC and ΔAIC using the starting Model (4) show outstanding performance to search for the true growth shape with the average hit rate of 97.1% (ranged between 96.8 and 97.7%) and 95.2% (ranged between 93.9 and 96.7%), respectively. As shown in **Table 2**, using ΔBIC for Model (4) shows consistently good performance regardless of other design factors, which are, true mean and covariance structure, covariate effect size, and sample size. Although LRT and ΔGFI (i.e., ΔCFI, ΔRMSEA, and ΔSRMR) show no advantages over the information criteria, when starting with Model (4), hit rates for ΔCFI and LRT notably increase with an average hit rate of 93.8 and 85.5%, respectively. In summary, ΔBIC and ΔAIC perform optimally to search for the accurate growth shape when starting with the most complex mean structure with the saturated error variance structure.

[5] Given the limited space, we present the summary of the simulation results. Result tables for six true model conditions including covariate effect size and sample size information are available from the first author upon request.

[6] Instead of the seventh-order polynomial model, we used the sixth-order polynomial model as the most complex mean model given that the seventh-order polynomial model resulted in serious nonconvergence issue.

TABLE 1 | Average percentage of finding the correct mean structure by four starting models.

						Linear growth			Quadratic growth		
Starting model[a]	**Cov spec**	**Mean spec**	**Effect size**	***n***	**Average hit**	**ID**	**UN(1)**	**AR(1)**	**ID**	**UN(1)**	**AR(1)**
(1)	ID	Step-up	0.1	100	56.0	76.1	65.6	51.1	64.0	49.8	29.3
(1)	ID	Step-up	0.1	210	48.1	80.9	57.5	31.7	65.1	36.7	16.8
(1)	ID	Step-up	0.1	390	41.2	83.9	45.5	14.2	65.3	25.6	12.6
(1)	ID	Step-up	0.3	100	57.7	78.3	68.3	51.5	64.4	53.8	29.9
(1)	ID	Step-up	0.3	210	50.2	81.6	61.6	32.2	65.3	42.1	18.1
(1)	ID	Step-up	0.3	390	43.3	84.5	50.7	14.4	65.1	30.9	14.2
(1)	ID	Step-up	0.5	100	58.6	79.4	70.1	51.9	64.9	53.5	31.8
(1)	ID	Step-up	0.5	210	51.1	82.6	65.1	32.8	65.4	40.9	20.0
(1)	ID	Step-up	0.5	390	44.6	86.1	56.2	14.6	65.1	29.3	16.0
Model (1) Average hit					50.1	81.5	60.1	32.7	64.9	40.3	21.0
(2)	ID	Top-down	0.1	100	56.6	71.3	62.2	49.3	86.3	41.8	28.7
(2)	ID	Top-down	0.1	210	50.7	76.5	55.6	31.0	91.7	32.5	16.7
(2)	ID	Top-down	0.1	390	45.5	80.0	44.9	14.1	93.9	27.7	12.2
(2)	ID	Top-down	0.3	100	58.5	73.6	64.8	49.6	86.7	45.7	30.5
(2)	ID	Top-down	0.3	210	52.7	77.2	59.7	31.4	91.9	35.9	20.2
(2)	ID	Top-down	0.3	390	47.6	80.9	50.1	14.3	94.1	30.4	15.8
(2)	ID	Top-down	0.5	100	60.4	74.7	66.6	50.1	87.5	49.1	34.6
(2)	ID	Top-down	0.5	210	55.2	78.5	63.2	31.9	92.1	40.1	25.6
(2)	ID	Top-down	0.5	390	50.7	83.0	55.5	14.4	94.0	35.9	21.6
Model (2) Average hit					53.1	77.3	58.1	31.8	90.9	37.7	22.9
(3)	UN	Step-up	0.1	100	73.8	81.0	84.1	83.3	66.7	69.1	58.7
(3)	UN	Step-up	0.1	210	83.4	88.3	89.1	91.5	79.5	80.6	71.6
(3)	UN	Step-up	0.1	390	86.3	90.6	91.5	93.7	83.6	83.8	74.4
(3)	UN	Step-up	0.3	100	64.2	85.7	86.4	88.3	43.0	44.6	37.4
(3)	UN	Step-up	0.3	210	73.4	88.8	89.8	91.7	58.2	59.0	52.6
(3)	UN	Step-up	0.3	390	77.9	91.4	92.4	93.8	64.6	64.3	60.6
(3)	UN	Step-up	0.5	100	52.0	86.7	87.4	89.2	16.8	16.9	15.1
(3)	UN	Step-up	0.5	210	58.1	89.8	90.8	91.9	25.5	26.8	23.6
(3)	UN	Step-up	0.5	390	65.8	92.3	93.1	94.0	38.6	39.9	36.6
Model (3) Average hit					70.5	88.3	89.4	90.8	52.9	53.9	47.8
(4)	UN	Top-down	0.1	100	76.6	67.8	70.7	71.2	82.8	83.8	83.3
(4)	UN	Top-down	0.1	210	82.5	75.0	75.9	80.8	87.2	88.1	87.7
(4)	UN	Top-down	0.1	390	85.3	78.7	79.8	84.9	89.4	89.3	89.7
(4)	UN	Top-down	0.3	100	78.8	72.4	72.9	76.2	83.2	84.2	83.8
(4)	UN	Top-down	0.3	210	82.9	75.7	76.8	81.1	87.5	88.2	87.9
(4)	UN	Top-down	0.3	390	85.8	79.8	81.0	85.2	89.4	89.4	89.7
(4)	UN	Top-down	0.5	100	79.5	73.2	73.8	77.0	83.7	84.6	84.4
(4)	UN	Top-down	0.5	210	83.5	76.9	77.9	81.6	87.8	88.5	88.1
(4)	UN	Top-down	0.5	390	86.2	81.0	81.9	85.7	89.5	89.5	89.7
Model (4) Average hit					82.3	75.6	76.7	80.4	86.7	87.3	87.1

[a] *Model (1): intercept-only with the simplest Identity V-CV structure, Model (2): highest-order polynomial growth (i.e., cubic for linear growth and sextic for quadratic growth population model) with the Identify V-CV, Model (3): intercept-only with the most complex UN V-CV structure, Model (4): highest-order polynomial growth with the UN V-CV structure.*

APPLIED STUDY

To illustrate the use of the proposed model search strategy, we have examined the longitudinal trajectories of depressive symptoms among Mexican American elders in the U.S. using the Hispanic Established Population for Epidemiological Studies of the Elderly (EPESE), which is retrieved from Inter-university Consortium for Political and Social Research (ICPSR). The first wave of interviews was conducted between September 1993 and June 1994 (Markides, 1993-1994), with 3,050 Mexican

TABLE 2 | Average percentage of finding the correct mean structure by different model evaluation criteria.

			Linear growth	Quadratic growth				
Selection criteria	**Starting model[a]**	**Overall average (%)**	**ID**	**UN(1)**	**AR(1)**	**ID**	**UN(1)**	**AR(1)**
LRT	(1)	53.1	95.3	65.2	29.3	94.4	28.9	5.6
	(2)	45.8	95.3	65.1	29.3	80.3	2.8	2.3
	(3)	85.6	94.9	95.1	95.0	76.6	78.3	73.9
	(4)	85.5	90.2	90.4	90.3	80.8	80.9	80.6
LRT Average		67.5	93.9	78.9	61.0	83.0	47.7	40.6
ΔCFI	(1)	20.9	75.1	33.4	16.9	0.0	0.0	0.0
	(2)	43.4	66.0	28.4	15.9	98.5	29.0	22.5
	(3)	56.3	83.0	86.2	94.9	33.6	23.8	16.4
	(4)	93.8	83.0	86.2	94.9	99.5	99.7	99.8
ΔCFI Average		53.6	76.7	58.5	55.6	57.9	38.1	34.7
ΔRMSEA	(1)	31.1	84.3	72.2	29.7	0.2	0.0	0.0
	(2)	55.0	74.6	67.5	26.2	82.9	62.1	16.6
	(3)	49.5	83.9	83.9	83.8	15.0	17.9	12.5
	(4)	57.4	56.4	56.3	56.6	58.4	58.3	58.5
ΔRMSEA Average		48.2	74.8	70.0	49.1	39.1	34.6	21.9
ΔSRMR	(1)	54.7	35.3	18.6	8.5	96.4	91.5	77.9
	(2)	47.9	30.8	17.4	8.1	89.3	81.6	59.9
	(3)	63.9	71.2	72.9	74.5	53.4	59.5	52.0
	(4)	63.0	28.9	30.8	45.0	89.3	92.4	91.4
ΔSRMR Average		57.4	41.5	34.9	34.0	82.1	81.3	70.3
ΔAIC	(1)	61.0	98.1	77.7	40.9	98.5	40.8	10.1
	(2)	53.7	96.7	76.6	40.3	94.3	8.3	6.4
	(3)	83.3	97.4	98.1	97.4	69.2	71.5	66.0
	(4)	95.2	96.0	96.7	96.0	93.9	94.3	94.1
ΔAIC Average		73.3	97.1	87.3	68.6	89.0	53.7	44.1
ΔBIC	(1)	83.5	99.0	98.4	81.2	100.0	86.8	35.8
	(2)	75.6	98.3	97.6	80.7	98.5	45.7	33.1
	(3)	81.2	97.6	98.6	97.7	64.8	67.5	61.1
	(4)	97.1	96.9	97.9	97.0	96.8	97.0	97.0
ΔBIC Average		84.4	97.9	98.1	89.1	90.0	74.3	56.7

[a] *Model (1): intercept-only with the simplest Identity V-CV structure, Model (2): highest-order polynomial growth (i.e., cubic for linear growth and sextic for quadratic growth population model) with the Identify V-CV, Model (3): intercept-only with the most complex UN V-CV structure, Model (4): highest-order polynomial growth with the UN V-CV structure.*

Americans aged 65 and over residing in the five southwestern states that contain the majority of Mexican Americans: Texas, California, New Mexico, Colorado and Arizona. Follow-up interviews were then conducted approximately every 2–3 years, with a supplemental sample from the same cohorts as the original sample added in wave 5. Literature has shown that limited English proficiency (LEP) is frequently reported to be associated with more depression among immigrants because language barriers can be a significant source of stress (Nwadiora and McAdoo, 1996; Constantine et al., 2004; Sadule-Rios, 2012). Kim et al. (in press) have investigated whether LEP is a significant factor associated with the longitudinal trajectory of the depressive symptoms using a latent growth model. In the current demonstration, we illustrate the model search procedure using the EPESE data to search for the optimal growth shape of the depressive symptoms for older immigrants.

Specifically, we have used a total of six waves of data for the depressive symptoms, which are measured with the Center for Epidemiologic Studies Depression Scale (CES-D), a 20-item self-administered questionnaire (Radloff, 1977), in the EPESE between 1993 and 2007. Respondents were asked to assess the frequency of depressive symptoms experienced during the past week, based on a 4-point scale with categories in the subsequent order: rarely or none of the time (0), some or a little of the time (1), much of the time (2), and most or all of the time (3). The total scores for 20 items potentially ranged from 0 to 60, with higher scores indicating more depressive symptoms. Among a total of 3,952 participants, 602 respondents who have all six waves

TABLE 3 | The AIC and BIC for unconditional growth models for EPESE data.

	Quartic	Cubic	Quadratic	Linear	Intercept
AIC	24,400	24,400	24,407	24,478	24,509
BIC	24,514	24,510	24,513	24,579	24,606

of data for CES-D are included in the further analysis. Mplus 7.3 using Maximum Likelihood estimation method (ESTIMATOR = MLR) was utilized for handling non-normality of the depression scores.

To search for the optimal growth trajectory, we analyzed a series of unconditional latent growth models (i.e., without having covariates) by changing the shape of the growth and compared the adjacent growth models using the information criteria (i.e., ΔBIC and ΔAIC), which showed the best hit rate for selecting the true population growth trajectory in simulations. Other fit indices (i.e., CFI, RMSEA, and SRMR) were also considered to meet the absolute fit criteria. First, based on our finding from the simulation study above, we specified the most complex (saturated) within-subject V-C structure (i.e., UN structure), which allows to freely estimate all the variance and covariance components. For the mean structure, we started with the quartic (i.e., 4th-order polynomial) growth model as the most complex mean structure with leaving one degree of freedom to generate the fit statistics for 6 waves of data[7]. Next, the quartic growth model with the UN error variance structure was compared with the cubic growth model, which has one less parameter in the mean structure to estimate. If there is no significant difference between the two competing models, we selected the simpler (cubic) growth model over the more complex (quartic) growth model. Next, the cubic growth model was compared to the quadratic growth model by eliminating the next highest-order polynomial growth term, and so on. When the model fit significantly got worse (i.e., ΔBIC > 2 and ΔAIC > 4), the model search was stopped and the more complex model was selected as the best fitting model.

Table 3 presents the model fit indices including the AIC and BIC for the series of latent growth models for the CES-D measures. As shown in **Table 3**, the cubic growth model was selected as the best fitting model by both information criteria. Using the ΔAIC, the quartic growth model shows no improvement from the cubic growth model, whereas the cubic growth model significantly better fits to the data than the quadratic growth model. Likewise, the cubic growth model is selected over the quartic growth model using the ΔBIC, and then, the cubic growth model is compared to the quadratic growth model, and indeed, the cubic growth model shows the better fit. Interestingly, the cubic growth model was selected by both step-up approach (i.e., starting from the intercept-only model) and top-down approach (i.e., starting from the quartic growth model) when specifying the saturated UN error variance structure in the current example. In other words, both starting points (i.e., simplest and the most complex) in terms of the mean structure reached to the same result in selecting the cubic growth model as the best fitting model. Results show that older immigrants' depressive symptoms have been decreased during the first two waves of data and then increased for the following four waves of data (see **Figure 3**). Further investigation and implication of the findings should be referred to the work by Kim et al. (in press).

DISCUSSIONS

The purpose of the current study is to explore the optimal model search strategy for searching for the best-fitting growth trajectory in latent growth models (LGM). While starting with the unconditional model without covariates has been known to be a classical method for model building in longitudinal data analysis (Meredith and Tisak, 1990; Singer and Willett, 2003), there is a lack of research incorporating both mean and residual variance structure in model search process under the framework of LGM. In the current study, we expanded the previous simulation study by Kim et al. (2016) by considering the time-invariant covariates on the growth trajectory to provide a model search strategy under more general conditions. We specifically examined two research questions: (a) which starting model performs best in searching for the correct growth trajectory, and (b) which model selection index performs best in identifying the true growth shape. Based on the results of the simulation study, we found that (a) starting with the *fully saturated model* with the most complex mean structure as well as the most relaxed (unstructured) error variance structure, and (b) using the information criteria (i.e., ΔBIC and ΔAIC) over the other fit evaluation criteria (i.e., LRT and ΔCFI, ΔRMSEA, and ΔSRMR) performed best in search for the population growth shape in LGM.

To examine the first research question, we have compared the four starting models in terms of the complexity of the mean structure and the within-subject variance-covariance (V-C) structure in LGM (**Figure 1**). For the within-subject V-CV structure, results of the simulations have shown that starting with the most complex (saturated) structure (i.e., Model 3 and 4) best recovers the true growth shape across all simulation conditions, which is a consistent finding with the previous simulation study (Kim et al., 2016). Unlike the previous study, however, the current results show that starting point in the mean structure also does matter to successfully search for the true growth trajectory. When there is a small to moderate effect of covariates regressed on growth trajectories, starting with the most complex mean structure outperforms the simplest mean structure to recover the true growth shape. This new finding is important because many applied research have been using the simpler starting model (e.g., intercept-only model or linear growth model) to search for the possibly more complex growth trajectory (e.g., quadratic growth or cubic growth model) in practice. Based on the simulation results, if the simpler growth model is used as the starting point, they are more likely to select the incorrectly simpler growth model, which may not represent their data adequately. As shown

[7] Fully saturated model in both mean and V-CV structures is just-identified model and generates no fit evaluation statistics other than the information criteria.

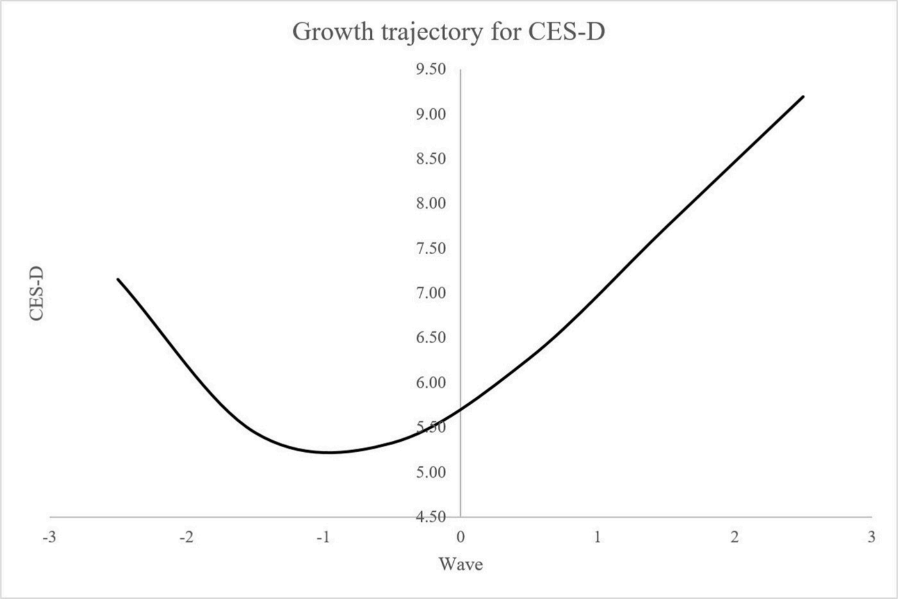


FIGURE 3 | Cubic growth trajectory for depressive symptoms using CES-D measure for EPESE data.

in the results of the simulation study (**Table 1**), when the true growth trajectory is a quadratic growth, the average hit rate of the fully saturated model (Model 4) is 87.0% while the hit rate of the simplest mean with the most complex error variance structure model (Model 3) is 51.5%. As sample size gets smaller and the covariate effect size gets larger, the impact of the starting point in the mean structure on recovering the true model becomes more substantial.

On the other hand, when the simplest ID structure is specified for the within-subject V-CV, neither step-up method nor top-down method successfully recovers the population growth trajectory except when the population model is generated to have the ID structure. Specifically, we note that the most commonly used intercept-only model with the simplest within-subject V-CV structure performs poorly to search for the correct mean trajectory, which is a consistent finding with the previous simulation study (Kim et al., 2016). When the true within-subject V-CV is not Identity but more complex structure (i.e., UN(1) or AR(1) in the current study), the average hit rate even decreases as sample size increases, which is another evidence of model instability. Given that researchers do not know the population or true variance structure in reality, specifying the simplest V-CV structure with no further consideration should be avoided in the model search process based on the current research finding.

To examine the second research question, we have used six model fit indices, which are LRT, ΔCFI, ΔRMSEA, ΔSRMR, ΔAIC, and ΔBIC to select for the best fitting growth trajectory model. Results show that there is no single fit index performing consistently well across all starting models. On the other hand, ΔBIC and ΔAIC performed successfully to search for the accurate growth trajectory with the use of the most complex starting model. As shown in the Appendix, average hit rates of ΔBIC and ΔAIC with using the Model (4) are above 95% on average across all simulation conditions. That being said, when researchers search for the optimal growth trajectory in LGM, starting with the most (or possibly more) complex mean structure with relaxing any constraints on error V-CV structure is highly recommended.

LIMITATIONS AND FUTURE

The current study has several limitations in study designs and conditions as with most simulation studies. First, we limited our study conditions for polynomial one-piece growth models (e.g., linear and quadratic growth models) in simulations based on the literature review, where majority of the applied research employed the polynomial growth models. While starting with a polynomial growth model is a reasonable approach, the proposed method might perform differently when the best model is a family of exponential growth models or piecewise growth models. Since the existence of multiple-piece non-linear model (e.g., piecewise exponential growth) is possible in reality, further research on the effectiveness of current approach with more complex multiple-piece models is needed. In addition, when the number of repeated measures is 3, this approach may not be adequate due to the limited number of testable growth models (intercept only, linear and quadratic).

Next, we used a single covariate with effect sizes to be equally regressed on all time factors (e.g., intercept and linear for a linear growth model). Some predictors may have a stronger effect on the initial time measure (e.g., intercept) than on the changing rate (e.g., linear and quadratic growth factors) or vice versa. Moreover, when there are multiple covariates or factors, models get easily complicated with possible interaction effects and effect sizes differed. We simplified the simulation conditions to use the constant effect sizes for the single covariate so that we can examine the effect of time-invariant covariates in model search process more clearly.

In the current study, we only considered the predictor(s) to be time-invariant covariates (e.g., gender, age, years of education, etc.) by excluding the scenarios for the time-varying covariates. Moreover, we have limited the assumption for the time-invariant covariates to be fully mediated by the growth parameters at the subject level. That is, we have assumed that the direct effect of time-invariant covariate on each time measure is equal to zero, which is regarded as a more standard way to model the time-invariant covariates in LGM (Whittaker and Khojasteh, 2017). While we believe that the current findings can be applied to more complex situations, further research is warranted to investigate the generalizability of the current research finding to more general conditions including the time-varying covariates in the population model.

This study has focused on the model specification search for finding for the accurate growth trajectory while having the search process for the residual variance structure left questionable. Given that the misspecified error variance structure has detrimental impacts on the inferences about growth parameters (Ferron et al., 2002; Kwok et al., 2007), searching for the correct or adequate error variance structure should be followed by specifying the optimal growth trajectory. Recently published simulation study by Ding et al. (2017) has provided a systematic approach to facilitate identifying a plausible covariance structure. Although they have conducted a study based on unconditional growth models, the guideline given in the study can be used as another starting point for searching the adequate error variance structure in LGM.

Implications and Practical Recommendations

Latent growth models are a popular method for longitudinal data analysis for decades given the flexibility of modeling the within- and between-subject error variance structure. This simulation study has investigated the performance of different starting models to search for the best-fitting growth trajectory in LGM under more general conditions than the previous simulation study. In the absence of certainty for the growth trajectory, the current study proposes to use the most complex (fully saturated) starting model with the highest-order polynomial growth factors and the most relaxed error variance structure, which performed the best to search for the true growth trajectory. Among the widely used fit indices for model comparisons (i.e., LRT, ΔGFI, and ΔIC), ΔBIC and ΔAIC with using the fully saturated starting model showed the most promising results in detecting the population growth trajectory over other fit indices. Based on the optimally specified growth trajectory, researchers should follow the next steps for model building process, such as, modeling the time-invariant and time-varying predictors, moderating effects, and specifying the proper covariance structures, to best understand the data and to examine their research questions.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2018.00349/full#supplementary-material

REFERENCES

Blozis, S. A. (2007). On fitting nonlinear latent curve models to multiple variables measured longitudinally. *Struct. Equat. Model. Multidisc. J.* 14, 179–201. doi: 10.1080/10705510709336743

Burnham, K. P., and Anderson, D. R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociol. Methods Res.* 33, 261–304. doi: 10.1177/0049124104268644

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Struct. Equat. Model.* 14, 464–504. doi: 10.1080/10705510701301834

Constantine, M. G., Okazaki, S., and Utsey, S. O. (2004). Self-concealment, social self-efficacy, acculturative stress, and depression in African, Asian, and Latin American international college students. *Am. J. Orthopsychiatry* 74:230. doi: 10.1037/0002-9432.74.3.230

Crockett, L. J., and Beal, S. J. (2012). The life course in the making: gender and the development of adolescents' expected timing of adult role transitions. *Dev. Psychol.* 48, 1727–1738. doi: 10.1037/a0027538

Ding, C. G., Jane, T.-D., Wu, C.-H., Lin, H.-R., and Shen, C.-K. (2017). A systematic approach for identifying level-1 error covariance structures in latent growth modeling. *Int. J. Behav. Dev.* 41, 444–455. doi: 10.1177/0165025416647800

Duncan, T. E., Duncan, S. C., and Stoolmiller, M. (1994). Modeling developmental processes using latent growth structural equation methodology. *Appl. Psychol. Meas.* 18, 343–354. doi: 10.1177/014662169401800405

Duncan, T. E., Duncan, S. C., and Strycker, L. A. (2013). *An Introduction to Latent Variable Growth Curve Modeling: Concepts, Issues, and Application*. Mahwah, NJ: Routledge Academic.

Ferron, J., Dailey, R., and Yi, Q. (2002). Effects of misspecifying the first-level error structure in two-level models of change. *Multivariate Behav. Res.* 37, 379–403. doi: 10.1207/S15327906MBR3703_4

Grimm, K. J., and Widaman, K. F. (2010). Residual structures in latent growth curve modeling. *Struct. Equat. Model.* 17, 424–442. doi: 10.1080/10705511.2010.489006

Hancock, G. R., and Lawrence, F. R. (2006). "'Using latent growth models to evaluate longitudinal change," in *Structural Equation Modeling: A Second Course*, eds G. R. Hancock and R. O. Mueller (Charlotte, NC: Information Age Publishing), 171–196.

Keijsers, L., and Poulin, F. (2013). Developmental changes in parent–child communication throughout adolescence. *Dev. Psychol.* 49, 2301–2308. doi: 10.1037/a0032217

Kim, G., Kim, M., Park, S., Jimenez, D. E., and Chiriboga, D. A. (in press). Limited English proficiency and trajectories of depressive symptoms among Mexican American elders. *Gerontologist*.

Kim, M., Kwok, O.-M., Yoon, M., Willson, V., and Lai, M. H. (2016). Specification search for identifying the correct mean trajectory in polynomial latent growth models. *J. Exp. Educ.* 84, 307–329. doi: 10.1080/00220973.2014.984831

Kwok, O.-M., West, S. G., and Green, S. B. (2007). The impact of misspecifying the within-subject covariance structure in multiwave longitudinal multilevel models: a monte carlo study. *Multivariate Behav. Res.* 42, 557–592. doi: 10.1080/00273170701540537

Leite, W. L., and Stapleton, L. M. (2011). Detecting growth shape misspecifications in latent growth models: an evaluation of fit indexes. *J. Exp. Educ.* 79, 361–381. doi: 10.1080/00220973.2010.509369

Liu, S., Rovine, M. J., and Molenaar, P. (2012). Selecting a linear mixed model for longitudinal data: repeated measures analysis of variance, covariance pattern model, and growth curve approaches. *Psychol. Methods* 17, 15–30. doi: 10.1037/a0026971

Markides, K. S. (1993-1994). *Data from: hispanic Established Populations for Epidemiologic Studies of the Elderly*. ICPSR02851-v2. Ann Arbor, MI: Inter-university Consortium for Political and Social Research.

Mayer, A., Steyer, R., and Mueller, H. (2012). A general approach to defining latent growth components. *Struct. Equat. Model. Multidisc. J.* 19, 513–533. doi: 10.1080/10705511.2012.713242

McCoach, D. B., and Kaniskan, B. (2010). Using time-varying covariates in multilevel growth models. *Front. Psychol.* 1:17. doi: 10.3389/fpsyg.2010.00017

Meredith, W., and Tisak, J. (1990). Latent curve analysis. *Psychometrika* 55, 107–122. doi: 10.1007/BF02294746

Muthén, L. K., and Muthén, B. O. (1998-2012). *Mplus User's Guide*. Los Angeles, CA: Muthén & Muthén.

Nwadiora, E., and McAdoo, H. (1996). Acculturative stress among Amerasian refugees: gender and racial differences. *Adolescence* 31, 477–488.

Preacher, K. J., Wichman, A. L., MacCallum, R. C., and Briggs, N. E. (2008). *Latent Growth Curve Modeling*. Los Angeles, CA: Sage.

Quené, H., and Van den Bergh, H. (2004). On multi-level modeling of data from repeated measures designs: a tutorial. *Speech Commun.* 43, 103–121. doi: 10.1016/j.specom.2004.02.004

Radloff, L. S. (1977). The CES-D scale: a self-report depression scale for research in the general population. *Appl. Psychol. Meas.* 1, 385–401. doi: 10.1177/014662167700100306

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociol. Methodol.* 111–163. doi: 10.2307/271063

Raudenbush, S. W., and Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods,* 2nd *Edn.* Thousand Oaks, CA: Sage Publications.

Raudenbush, S. W., and Xiao-Feng, L. (2001). Effects of study duration, frequency of observation, and sample size on power in studies of group differences in polynomial change. *Psychol. Methods* 6, 387–401. doi: 10.1037/1082-989X.6.4.387

Ryoo, J. H. (2011). Model selection with the linear mixed model for longitudinal data. *Multivariate Behav. Res.* 46, 598–624. doi: 10.1080/00273171.2011.589264

Sadule-Rios, N. (2012). A review of the literature about depression in late life among Hispanics in the United States. *Issues Ment. Health Nurs.* 33, 458–468. doi: 10.3109/01612840.2012.675415

Singer, J. D., and Willett, J. B. (2003). *Applied Longitudinal Data Analysis: Modeling Change And Event Occurrence*. New York, NY: Oxford university press.

Titzmann, P. F., Silbereisen, R. K., and Mesch, G. (2014). Minor delinquency and immigration: a longitudinal study among male adolescents. *Dev. Psychol.* 50, 271–282. doi: 10.1037/a0032666

Van der Graaff, J., Branje, S., De Wied, M., Hawk, S., Van Lier, P., and Meeus, W. (2014). Perspective taking and empathic concern in adolescence: gender differences in developmental changes. *Dev. Psychol.* 50, 881–888. doi: 10.1037/a0034325

Whittaker, T. A., and Khojasteh, J. (2017). Detecting appropriate trajectories of growth in latent growth models: the performance of information-based criteria. *J. Exp. Educ.* 85, 215–230. doi: 10.1080/00220973.2015.1123669

Wu, W., and West, S. G. (2010). Sensitivity of fit indices to misspecification in growth curve models. *Multivariate Behav. Res.* 45, 420–452. doi: 10.1080/00273171.2010.483378

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 22 February 2018
doi: 10.3389/fpsyg.2018.00130

Evaluation of Analysis Approaches for Latent Class Analysis with Auxiliary Linear Growth Model

Akihito Kamata[1]*, *Yusuf Kara*[2], *Chalie Patarapichayatham*[3] and *Patrick Lan*[3]

[1] Department of Psychology, Department of Education Policy and Leadership, Center on Research and Evaluation, Southern Methodist University, Dallas, TX, United States, [2] Department of Educational Measurement and Evaluation, Anadolu University, Eskisehir, Turkey, [3] Simmons School of Education, Southern Methodist University, Dallas, TX, United States

This study investigated the performance of three selected approaches to estimating a two-phase mixture model, where the first phase was a two-class latent class analysis model and the second phase was a linear growth model with four time points. The three evaluated methods were (a) one-step approach, (b) three-step approach, and (c) case-weight approach. As a result, some important results were demonstrated. First, the case-weight and three-step approaches demonstrated higher convergence rate than the one-step approach. Second, it was revealed that case-weight and three-step approaches generally did better in correct model selection than the one-step approach. Third, it was revealed that parameters were similarly recovered well by all three approaches for the larger class. However, parameter recovery for the smaller class differed between the three approaches. For example, the case-weight approach produced constantly lower empirical standard errors. However, the estimated standard errors were substantially underestimated by the case-weight and three-step approaches when class separation was low. Also, bias was substantially higher for the case-weight approach than the other two approaches.

Keywords: mixture model, latent class analysis, case-weight approach, one-step approach, three-step approach

OPEN ACCESS

Edited by:
Oi-Man Kwok,
Texas A&M University, United States

Reviewed by:
Eun Sook Kim,
University of South Florida, United States
Minjung Kim,
The Ohio State University, United States

****Correspondence:***
Akihito Kamata
akamata@smu.edu

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *30 March 2017*
Accepted: *26 January 2018*
Published: *22 February 2018*

Citation:
Kamata A, Kara Y, Patarapichayatham C and Lan P (2018) Evaluation of Analysis Approaches for Latent Class Analysis with Auxiliary Linear Growth Model. Front. Psychol. 9:130. doi: 10.3389/fpsyg.2018.00130

INTRODUCTION

Mixture modeling has become a widely used statistical method in behavioral sciences because it allows for an exploration of identification and understanding of latent subpopulations in a given population. Among them, a method where categorical latent trait constructs are identified based on multiple observed categorical variables is specifically referred to as a latent class analysis (LCA) (Lazarsfeld and Henry, 1968; Dayton and Macready, 1998). While identifying and interpreting latent classes may be of the main interest with LCA, researchers may be also interested in how the identified latent classes are related to auxiliary variables, such as covariates and distal outcomes. In other words, researchers are not only interested in latent classes of individuals, but also in potential causes and/or consequences of the class membership (Bakk et al., 2013, 2014). This type of analysis would provide additional information about heterogeneity of the relations, since it is not realistic to assume that all individuals in the population have the same relations to auxiliary variables (Nylund-Gibson et al., 2014). Moreover, researchers may be interested in considering an auxiliary model in conjunction with LCA, such that separate auxiliary model parameters are estimated for each of the latent classes. For example, a simple linear regression model as an auxiliary model to LCA was presented and investigated in Asparouhov and Muthén (2014). In such a modeling, the latent class variable can be thought of a moderator for the auxiliary model (i.e., secondary model). In this paper, this type of a model is referred to as a two-phase mixture model, because the model is consisted of two phases, the LCA model phase and the auxiliary model phase.

One may argue that a single mixture model without a latent class measurement model may be sufficient to describe heterogeneity on the auxiliary model, such as mixture regression and growth mixture model. However, there are contexts where latent classes should be defined by a latent class measurement model, rather than by a single mixture model. For example, Asparouhov and Muthén (2014) and Vermunt (2010) pointed out that a single mixture model approach will not fit a logic of a researcher, if the latent class measurement model is theorized to define latent classes, rather than the mixture distribution of the auxiliary variable or model. In such a case, results from the two-phase mixture model are not necessarily the same as results from the single mixture model. Therefore, it is paramount to identify latent classes by measurement indicators in the latent class measurement model first, rather than directly attempting to identify latent classes based on heterogeneity in their auxiliary variable or model. Thus, an implementation of a two-phase mixture model becomes important.

METHODS TO TWO-PHASE MIXTURE MODELS

There are several different approaches that can be undertaken to estimate a two-phase mixture model. In this section, four selected approaches are described, although the first approach will not be investigated in this study.

Classify and Analyze Approach

Classify-and-analyze approach is a two-step process, also referred to as hard partitioning (Vermunt, 2010). In the first step, LCA is conducted, and each individual is assigned to a specific latent class by the highest posterior class-membership probability that is obtained from the LCA. Then, in the second step, class assignments are used as an observed grouping variable to compare groups on auxiliary variables, if the model contains auxiliary variables. If the model contains auxiliary model, the auxiliary model will be fitted for each of identified classes. In either case, membership in identified classes is mutually exclusive, such that each observation is classified into only one of the identified classes. While it is straightforward to implement (Hibbard et al., 2007; Reinke et al., 2008; Archambault et al., 2009; Hardigan, 2009), this strategy comes with some critical disadvantages. First, there can be misclassified individuals, because deterministic classifications are based on the probabilistic information of class-membership probabilities. It is known that misclassification of individuals in the classify-and-analyze approach can result in biased estimates of the relations between the latent classes and the auxiliary variables and auxiliary model parameters (Hagenaars, 1993; Clogg, 1995). Second, somewhat related to the first disadvantage, classification uncertainties (namely, measurement errors in classifications from the LCA) would be ignored. Since classifications are treated as true states, the standard errors for parameter estimates by the classify-and-analyze approach are likely underestimated (Roeder et al., 1999; Loken, 2004; Clark and Muthén, 2009). Overall, the literature to date is in agreement that the classify-and-analyze approach is no longer recommended for estimating an LCA model with auxiliary variables and/or auxiliary models. Therefore, the classify-and-analyze approach was not considered further in this study.

One-Step Approach

The one-step approach involves a simultaneous estimation of an LCA model and auxiliary variables and/or auxiliary models (Formann, 1992; Heijden et al., 1996; Bandeen-Roche et al., 1997; Dayton and Macready, 1998; Muthén and Muthén, 2000; Clark and Muthén, 2009; Kim et al., 2016). The one-step approach is recommended particularly by earlier literature (Heijden et al., 1996; Muthén, 2001), because estimating LCA and auxiliary models in one-step has advantages over the classify-and-analyze approach. First, occurrence of classifying individuals into incorrect classes would be irrelevant, because the one-step approach does not involve classifications of individuals into particular classes based on estimated class probabilities. In other words, the estimation of the latent classes is accomplished jointly by the inclusion of auxiliary variable(s) and/or model(s) (Kim et al., 2016). As underlined by Clark and Muthén (2009), individuals can be fractional members of all identified latent classes in the one-step approach. Thus, it reduces problems that arise from treating the latent classes as a true state, the procedure that is followed by the classify-and-analyze approach. Second, measurement errors of class membership would be incorporated in the analysis, because they are embedded in the model by the one-step approach. Another advantage of the one-step approach is a contribution of the included auxiliary variable(s)/model to the estimation of latent classes. Clark and Muthén (2009) argue that this inclusion improves the class separation and reduces the standard errors.

However, while it is still known as an efficient approach, recent studies are cautious about employing the one-step approach (Vermunt, 2010; Nylund-Gibson et al., 2014). The prominent reason is that the parameters of the first-phase LCA model may be affected by auxiliary variables and/or models, if the strength of the associations between latent class indicators and latent classes are not sufficiently strong (Vermunt, 2010; Asparouhov and Muthén, 2014). If this becomes a problem, it could lead to a different number and/or interpretations of latent classes by including auxiliary variables and/or models. Changing the parameters in this manner would be disconcerting and leads to problems with model construction. While the inclusion of auxiliary variables and/or models is important, the measurement of the latent classes should be free from influence of auxiliary variables and models (Nylund-Gibson et al., 2014).

Three-Step Approach

Another approach to a two-phase mixture model is the three-step approach (Bolck et al., 2004; Vermunt, 2010). The key advantage of the three-step approach is a separate treatment of the LCA model and auxiliary variables or models, just like classify-and-analyze approach, while classification measurement errors are still taken into account. As a result, class separation is accomplished without being affected by auxiliary variables and models (Vermunt, 2010; Kim et al., 2016). As the first step with the three-step approach, the LCA model is estimated as a measurement model by using only latent class indicator variables. In the second step, a variable for most likely classes ($\aleph$) is created by the modal assignment using the largest posterior probabilities obtained in the first step. Just like classify-and-analyze approach, $\aleph$ is treated as a manifest nominal variable that represents the class assignments. However, the three-step approach retains the information about classification uncertainties and utilizes it as the measurement errors of classifications as follows. Using the estimated posterior class probabilities and number of the individuals assigned to each of the latent classes, classification uncertainty rates are computed. These rates are the average posterior probabilities in the form of $k \times k$ matrix, where k is the number of latent classes. In the third step, the auxiliary model is fit separately for each of the identified classes in the first step by incorporating the measurement errors derived in the second step. Bolck et al. (2004) demonstrated their three-step approach underestimated associations between class membership and auxiliary variables. Vermunt (2010) proposed a correction method by maximizing a weighted log-likelihood function for clustered data. With a series of simulation studies, Vermunt demonstrated that the correction improved the method substantially. Currently, the three-step approach with Vermunt's correction is incorporated in Mplus software (Asparouhov and Muthén, 2014).

Asparouhov and Muthén (2014) demonstrated that the three-step approach with Vermunt's correction recovered parameters very well, when the latent class variable was measured well by the LCA model (i.e., high entropy). Also, it was demonstrated that the loss of efficiency for the three-step approach was minimal, compared to the one-step approach. On the other hand, Bakk et al. (2014) reported that the bias-corrected three-step approach utilized in Mplus software tends to underestimate the standard errors of the auxiliary variables effects. Nylund-Gibson et al. (2014) extended the application of this three-step approach to a latent transition analysis (LTA). Overall, the three-step approach with Vermunt's correction has become a promising method to estimate a mixture model with auxiliary variables and/or auxiliary models. Nonetheless, Asparouhov and Muthén (2014) argued that any method could fail to achieve satisfactory accuracy and efficiency, if the latent class variable is poorly measured by the measurement model (i.e., low entropy), including the three-step approach.

Case-Weight Approach

The case-weight approach for mixture models is also a three-step procedure. In the first step, the measurement model (i.e., LCA) is estimated by using only latent class indicator variables. In fact, this first step LCA is exactly the same as the first step of the aforementioned three-step approach. However, how the information about classification uncertainties are derived in the second step is different from the three-step approach. In the second step of the case-weight approach, the estimated posterior class probabilities from the first step are directly saved as weight variables (one weight variable for each identified class). In the third step, the auxiliary model is fit separately for each of the latent classes by using the corresponding weight variable from the second step as the case weights.

This way, each observation is treated as a fractional member of all identified latent classes, as a way to incorporate classification uncertainties. As a result, the contribution of each observation to a given class is represented by the estimated class probability for the observation. For example, if an observation has a very small class probability for a given latent class, the observation will have a very small impact on estimating parameters of an auxiliary model, but not zero. Also, the effective sample size for each class is the sum of the estimated class probabilities, which is a reasonable realization of the estimated class size. This procedure is analogous to computing a weighted data summary quantity, such as a weighted mean, which is also similar to the propensity score weighting procedure (Robins and Rotnitzky, 1995; Hirano and Imbens, 2001).

As one example related to this approach, Clark and Muthén (2009) demonstrated an approach, where the latent class variable was regressed on a predictor variable by using the classification probabilities from the initial-step LCA as regression weights. Cheng (2012) also employed the same approach for an LCA model with a distal outcome. Clark and Muthén, as well as Cheng, confirmed that the weighted regression approach worked well, while the one-step approach was still found to best account for the uncertainty in latent class membership.

The case-weight approach discussed in this paper assumes any kind of latent-class measurement model and any kind of auxiliary model. For example, Nese et al. (2017) employed this approach to study heterogeneity of the growth of emergent literacy knowledge by combining a zero-inflated Poisson regression model (i.e., the latent-class measurement model phase) and a three-class growth mixture model (i.e., the auxiliary model phase). However, the performance of this approach is rather unknown. Thus, the current study aimed to investigate the performance of the case-weight approach under various conditions for a two-phase mixture model through a simulation study. The performance of the case-weight approach was also compared to two other approaches; namely, one-step and three-step approaches.

METHODS

Model

The first phase of the investigated two-phase mixture model was a two-class LCA model with four dichotomous measurement indicators. The model is expressed as

$$P\left(U_p = 1|c\right) = \left[1 + \exp\left(\tau_{cp}\right)\right]^{-1},$$

where U_p is the response on the pth dichotomous measurement indicator ($p = 1, \ldots, 4$) and c is the latent class variable ($c = 1$ or 2). Also, τ_{cp} is the threshold parameter for pth measurement indicator for latent class c. Accordingly, τ_{cp} is the logit of $U_p = 1$, given in the cth class.

The second phase of the two-phase mixture model was an auxiliary model, which was a linear growth model (LGM) with four time points. The LGM was set up as a special case of a two-factor confirmatory factor analysis model, where the two latent factors represented the growth intercept and growth slope that varied between individuals. The model is expressed as

$$\mathbf{y} = \boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\epsilon},$$

where $\mathbf{y}$ is a 4 × 1 vector of outcome measures, $\boldsymbol{\Lambda}$ is a 4 × 2 matrix of factor loadings, $\boldsymbol{\eta}$ is 2 × 1 vector of two latent factors, and $\boldsymbol{\epsilon}$ is a 4 × 1 vector of residuals. Factor loadings for the four outcome measures were all constrained to fixed values: [1, 1, 1, 1] for the intercept factor (the first column of $\boldsymbol{\Lambda}$), and [0, 1, 2, 3] for the slope factor (the second column of $\boldsymbol{\Lambda}$). As a result, the growth intercept was a realization of the initial status. In addition, $\boldsymbol{\epsilon}$ was assumed to be normally distributed with 0 means and covariance matrix with equal diagonals and 0 off-diagonals, indicating that error variances for the four outcome measures were constrained to be equal and zero covariances between errors. In addition, $\boldsymbol{\eta}$ was assumed to be normally distributed with unknown means (mean intercept and mean growth trajectory) and covariance matrix (variances of intercept and growth trajectory, and covariance between intercept and growth trajectory). All parameters in the auxiliary model (i.e., mean intercept, mean slope, intercept variance, slope variance, covariance between intercept and slope, and error variances) were assumed to be different between the two latent classes. A graphical representation of this two-phase mixture model is also provided in **Figure 1**. As mentioned above, all parameters in the auxiliary model were assumed to be different between latent classes. These parameters are graphically indicated as dots on straight and curved arrows in **Figure 1**.

The true parameter values for the LCA model were varied, including the class proportion for the smaller class. Hereafter, the smaller class will be referred to as "class 2." The threshold parameters were constrained to be the same for the four measurement indicators, but the value was varied depending on simulation conditions (see below). These differences in threshold parameter values indirectly affected differences in class separation (i.e., entropy), where a lower threshold resulted into a lower class separation. Parameters for the auxiliary LGM were assumed to be different between the two classes, but fixed for all simulation conditions. The parameter values for the auxiliary model are provided in **Table 1**. Note that we did not hypothesize any direct relations between the auxiliary model variables and latent class indicators, just like in Bakk et al. (2013).

Simulation Study

Data were generated for the two model phases simultaneously, just like how Asparouhov and Muthén (2014) generated data. According to Asparouhov and Muthén, this data generation strategy generates data that would be consistent with a 2-phase mixture model, because the latent class variable is not an endogenous variable in the data generation model. Data sets were generated for a total of 27 within-method simulation conditions, with a minimum of 1,000 replications for each condition. We generated additional replications if there were fewer than 1,000

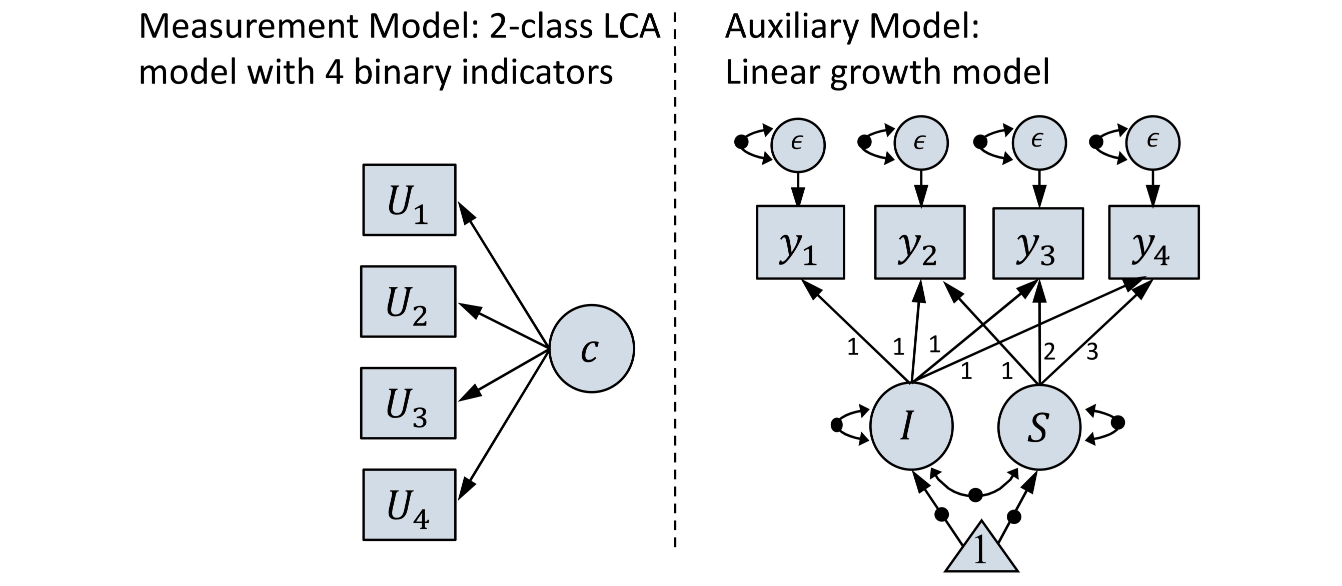

Measurement Model: 2-class LCA model with 4 binary indicators
U_1
U_2
U_3
U_4
c
Auxiliary Model: Linear growth model
ϵ
ϵ
ϵ
ϵ
y_1
y_2
y_3
y_4
1
1
1
1
1
2
3
I
S
1

FIGURE 1 | Graphical representation of the studied two-phase mixture model. On the measurement model, U_1, U_2, U_3, and U_4 represent four dichotomous outcome variables related to latent class variable c. On the auxiliary model, y_1, y_2, y_3, and y_4 represent repeatedly measured outcome variable at four time points. Also, I represents the growth intercept, S represents the growth slope, and ϵ represent the residuals. Black dots indicate that parameters represented by these arrows are different between latent classes.

TABLE 1 | True parameter values for the auxiliary model.

Parameter	Class 1 (larger class)	Class 2 (smaller class)
Mean(*I*)	0.6	0.4
Mean(*S*)	1.0	1.8
Variance(*I*)	1.9	1.4
Variance(*S*)	0.4	0.3
Covariance(*I*, *S*)	0.5	0.3
Variance(ϵ)	0.5	0.7

I, intercept; S, slope; ϵ, residuals.

successfully converged replications that correctly identified the 2-class model as the best model by BIC for any of the analysis methods. We followed this strategy only for fitting the 2-class model, because the parameter recovery evaluations were undertaken only when the 2-class model was fitted. In addition, if any methods that had more than 1,000 successfully converged replications with 2-class model as the best model by BIC for a particular condition, only the first 1,000 replications were evaluated for parameter recovery evaluations.

The 27 within-method simulation conditions were represented by three simulation factors; namely, sample sizes, class proportion for the smaller class (class 2), and class separation (i.e., threshold parameter in the LCA phase of the model). These three simulation factors were chosen, because they are known to affect the performance of mixture model estimation. Three sample sizes were: small (500 examinees), medium (1,000 examinees), and large (2,000 examinees). Three class-2 proportions were: small (0.05), medium (0.15), and large (0.30). Note that this study generated latent classes only by a two-class LCA model. Lastly, three levels of class separation (the threshold parameter the LCA phase of the model) were: low (0.754), medium (1.254), and high (1.750). These threshold parameter values were computed by first defining the log-odds difference between classes for the LCA phase of the model; low = 1.50, medium = 2.50, and high = 3.50. As a result, the average entropy was 0.66, 0.77, and 0.90 for the three levels of the class separation in the simulation. Data generated for each of the 27 within-method simulation conditions were fitted by three methods, namely, one-step approach (OS), case-weight approach (CW), and three-step approach (TS).

For each simulation condition, the model fit for the 2-class model was evaluated relative to 1-class and 3-class models. To do so, the proportion of replications, in which Bayesian information criterion (BIC) for the 2-class model was smaller than ones for 1-class and 3-class models, was computed for each of the three methods for each of the 27 within-method simulation conditions. For the case-weight and three-step approaches, this evaluation was commonly performed for the first-step LCA model, because it would be the step where one would make a model selection decision regarding the number of latent classes for these two approaches. Also, convergence rate was evaluated for the 1-class, 2-class, and 3-class models for each simulation condition. Note that a computation of the convergence rate for the CW and TS approaches involved a multiplication of the convergence rate of the first-step LCA model and the convergence rate of the third-step auxiliary LGM model.

Finally, parameter recovery performance was evaluated for the 2-class model, separately for the three approaches for each auxiliary model parameter for the two latent classes for each of the 27 within-method simulation conditions, by computing; (a) absolute relative bias, (b) empirical standard error (SE), (c) the mean estimated SE relative to the empirical SE, and (d) root mean square error (RMSE). Then, each of the four indices were averaged across all model parameters for the two latent classes separately for each of the 27 within-method simulation conditions. As mentioned earlier, only the first 1,000 successfully converged replications were included in the parameter recovery evaluations, including only replications that concluded the 2-class model was correctly selected by the BIC.

Note that a bias is the systematic part of the estimation error. In this study an absolute relative bias was computed by taking the absolute value of a relative bias value (i.e., bias divided by the true parameter value). For a given parameter θ ,

$$(absolute\ relative\ bias)_\theta = \left| \frac{\left(\frac{\sum_{i=1}^{r} \widehat{\theta}_i}{r} \right) - \theta}{\theta} \right|,$$

where $\hat{\theta}_i$ is the parameter estimate for the *i*th replication, θ is the true parameter value, and r is the number of replications. On the other hand, an empirical SE is the random part of estimation error that attributes to sampling and was computed as the standard deviation of repeatedly obtained 1,000 parameter estimates for a given parameter θ by

$$\left(empirical\ \ SE\right)_\theta = \sqrt{\frac{\sum_{i=1}^{r} \left(\widehat{\theta}_i - \left(\frac{\sum_{i=1}^{r} \widehat{\theta}_i}{r} \right) \right)^2}{r}},$$

where all symbols are defined above. Also, each simulation replication produced an estimated SE, and it is explicitly referred to as the "estimated SE" in this study to distinguish it from the empirical SE. The empirical SE is a numerically realized theoretical SE based on repeatedly sampled data, while the estimated SE is an analytically (or numerically, in some other cases, such as the bootstrap method) estimated SE based on one given sample data. In practice, only an estimated SE will be available to data analysts and will be treated as the best estimate of the theoretical SE. Therefore, it would be of interest how much the estimated SEs are close to the theoretical SE (i.e., the empirical SE) to evaluate the quality of the estimated SEs. Therefore, the mean of the estimated SEs was computed across 1,000 replications, and its magnitude was compared to the empirical SE by their ratio to evaluate potential under- or over-estimation of the estimated SEs. Finally, RMSE is the total estimation error, and it was computed for a given parameter θ by

$$(RMSE)_\theta = \sqrt{\frac{\sum_{i=1}^{r} \left(\widehat{\theta}_i - \theta \right)^2}{r}},$$

where all symbols are defined above.

Mplus software (Muthén and Muthén, 1998–2012) was used to generate the data, as well as to fit the model. Data generations and analyses with Mplus were controlled by R software (R Core Team, 2016). Examples of Mplus syntax are provided as a Supplementary Material.

RESULTS

Convergence Rate

Convergence rates are summarized in **Table 2**. Although they are not shown in the table, all replications converged without any warning or error for the 1-class one-step approach and first step 1-class LCA model. Also, almost all replications of the first-step 2-class LCA model converged, which was shared by the case-weight and three-step approaches, with the lowest convergence rate of 97.1%.

For the 2-class model, the case-weight approach had the highest convergence rate among the three methods. For example, they converged nearly 100% for all conditions when $n = 2{,}000$, while its convergence rate dropped somewhat when the class-2 proportion was small with $n = 500$. Nonetheless, its convergence rates were always higher than 96%. The convergence rates for the three-step approach had a similar pattern as the case-weight approach, namely, when class-2 proportion was small, convergence rate was lower. However, the convergence rates were constantly lower than the ones for the case-weight approach within the same conditions. In some conditions, they were substantially lower, especially when $n = 500$, and/or when the class-2 proportion was small. Even with $n = 2{,}000$, when the class-2 proportion was small and the class separation was low, the convergence rate dropped to 56.9%, whereas the convergence rate remained nearly 100% for the case-weight approach. On the other hand, the convergence rate for the one-step approach dropped to even lower percentages with lower sample size, smaller class-2 proportion, and/or lower class separation. For example, the convergence rate was 79.6% when the class-2 proportion was small and the class separation was low even with $n = 2{,}000$. It dropped to only 29.4% in the same condition with $n = 500$.

TABLE 2 | Percentages of convergence and correct model selection.

Sample size	Class-2 proportion	Class separation	Convergence: OS approach		Convergence: CW approach		Convergence: TS approach		Correct model selection	
			2-Class	3-Class	2-Class	3-Class	2-Class	3-Class	OS	LCA
$n = 500$	Small	Low	29.4	2.7	96.9	80.1	50.0	32.7	11.3	7.1
		Medium	64.5	5.8	98.9	72.7	71.1	37.0	61.3	86.0
		High	80.5	7.3	97.5	62.4	81.2	38.7	79.7	99.5
	Medium	Low	83.3	6.9	99.4	78.3	84.5	40.1	81.2	70.4
		Medium	94.8	9.1	100.0	53.8	94.5	38.1	93.9	99.8
		High	98.4	9.3	100.0	51.4	97.9	42.1	97.9	99.8
	Large	Low	98.3	8.8	100.0	75.5	97.4	49.9	97.2	97.8
		Medium	99.9	11.7	100.0	47.5	99.5	41.2	98.9	100.0
		High	99.9	10.5	100.0	53.6	99.9	48.3	98.6	99.6
$n = 1{,}000$	Small	Low	53.2	5.2	96.8	74.8	51.4	26.3	24.2	9.2
		Medium	84.6	9.4	99.7	62.9	85.5	27.4	84.5	98.1
		High	93.7	11.9	99.8	48.0	94.6	31.4	93.7	100.0
	Medium	Low	96.1	10.4	99.9	67.8	94.6	33.9	96.0	94.2
		Medium	99.4	11.3	100.0	41.9	99.2	28.8	99.4	100.0
		High	99.9	11.2	100.0	36.0	99.9	25.9	99.9	100.0
	Large	Low	99.9	11.9	100.0	55.5	99.9	35.0	99.8	99.7
		Medium	100.0	11.2	100.0	35.3	100.0	29.2	99.9	100.0
		High	100.0	12.3	100.0	44.7	100.0	37.7	100.0	100.0
$n = 2{,}000$	Small	Low	79.6	8.8	98.0	69.1	56.9	15.3	61.2	19.8
		Medium	96.2	12.3	100.0	53.1	95.1	20.5	96.2	100.0
		High	98.6	15.3	100.0	38.2	98.7	19.4	98.6	100.0
	Medium	Low	99.7	11.3	100.0	55.5	99.3	20.3	99.7	100.0
		Medium	100.0	12.8	100.0	31.6	100.0	18.1	100.0	100.0
		High	100.0	12.7	100.0	33.9	100.0	20.9	100.0	100.0
	Large	Low	100.0	11.8	100.0	36.7	100.0	22.9	100.0	100.0
		Medium	100.0	12.8	100.0	31.2	100.0	23.3	100.0	100.0
		High	100.0	15.3	100.0	40.7	100.0	31.9	100.0	100.0

OS, one-step approach; CW, case-weight approach, and TS, three-step approach. LCA was common first step for CW and TS approaches.

For the 3-class model, convergence rates for the one-step approach dramatically dropped. The highest convergence rate was only 15.3% for the conditions with $n = 2{,}000$ and high class separation. On the other hand, the convergence rates remained high for the case-weight approach, although they were uniformly lower than 2-class model in comparable conditions. For the three-step approach, convergence rates for 3-class model dropped much more than the case-weight approach. Yet, convergence rates were considerably higher than the ones for the one-step approach.

Model Selection

Percentages of correct model selection are summarized in the last two columns of **Table 2**. First, correct model selection rates were quite low either by the one-step approach or the first-step LCA when class separation was low and class-2 proportion was small. For this combination of the conditions, correct model selection rates were always low, regardless of the sample size.

On the other hand, correct model selection rates were 100% or nearly 100% with high class separation and large or medium class-2 proportion, regardless of the analysis method and the sample size. Also, conditions with medium class separation and large class-2 proportion demonstrated quite high correct model selection rates. With high class separation and small class-2 proportion, the correct model selection rate was nearly 100% with $n = 2{,}000$ (98.6% for OS and 100% for first-step LCA). However, the rates decreased as the sample size became smaller for OS; 93.7% with $n = 1{,}000$, and 79.7% with $n = 500$, while the rates remained near 100% for the first-step LCA. Similar patterns were observed for conditions with medium class separation and medium class-2 proportion.

Overall, the first-step LCA (i.e., case-weight approach and three-step approach) was better in correct model selection than the one-step approach. Exceptions were when class separation was low and class-2 proportion was small. Another exception was when class separation was low and class-2 proportion was medium with $n = 500$.

Parameter Recovery

As mentioned earlier, parameter recovery results were summarized by averaging for all parameters in the auxiliary LGM for each latent class. The summary results are presented in **Figure 2** (mean of absolute relative bias), **Figure 3** (mean of empirical SE), **Figure 4** (mean of estimated SE relative to empirical SE), and **Figure 5** (mean of RMSE). For each figure, results are summarized into three columns of graphs for three sample sizes ($n = 500$; $n = 1{,}000$; $n = 2{,}000$) for each latent class. The first three columns of graphs are for the larger class (class 1), and the last three columns of graphs are for class 2 (smaller class). Three rows of graphs are for the three levels of the class-2 proportion (small; medium; large). The three ticks on the horizontal axis of each graph are three levels of class separation (low; medium; high).

For the larger class (class 1), all of absolute relative bias, empirical SE, and RMSE were substantially smaller. Particularly, differences between the three approaches were nearly undistinguishable for class 1 for high class-separation conditions, regardless of sample size and class-2 proportion. The only exception was the relative estimated SE, where underestimation of the estimated SE was revealed for the case-weight and three-step approaches, especially when class separation was low. Underestimation of estimated SE was nearly zero for conditions with medium or high class separation for all three approaches. Interestingly, underestimation was much larger by the one-step approach than the other two approaches when class separation was low, class-2 proportion was small, and $n = 500$.

There were some important observations for results for the smaller class (class 2). Hereafter, discussions of the results are focused on class 2. First, it was revealed that the mean of absolute relative bias (**Figure 2**) was larger for the case-weight approach than the other two approaches in all conditions. Relative bias for the one-step approach and the three-step approach sharply decreased as the sample size became larger, as the class separation became higher, and as the class-2 proportion became larger. However, relative bias for the case-weight approach was affected much less by the class-2 proportion and the sample size, while it was still affected by the class separation. In other words, larger sample size and larger class-2 proportion did not reduce the relative bias by the case-weight approach. On the other hand, relative bias for all three approaches decreased sharply as the class separation became higher, and the discrepancy between the case-weight approach and the other two approaches became smaller when the class separation was high. Overall, the one-step and three-step approaches displayed strength with respect to relative bias, while the case-weight approach did not.

Although details are not presented in this paper, results for each parameter were examined under $n = 500$ conditions. The mean and variance parameters of the slope for class 2 was particularly high in relative bias by all three approaches when the class-2 proportion was small and the class separation was low. However, sharp decrease was observed for all three approaches as the class separation became higher. Also, sharp decrease was observed for the one-step and three-step approaches as the class-2 proportion became larger. Overall, it was confirmed that relative bias for the case-weight approach was constantly higher than the two other approaches for all parameters for the smaller class. Also, it was confirmed that the discrepancy between the three approaches became smaller as the class separation became higher.

With respect to empirical SE (**Figure 3**), the performance of the case-weight approach was better than the other two approaches, especially when the class-2 proportion and the sample size was small. However, the discrepancies between the three approaches became smaller as the class separation became higher and the class-2 proportion became larger. The performance of the one-step and three-step approaches were similar; when the class separations were medium or high, their empirical SEs were nearly identical, especially under medium and large class-2 proportion conditions. Overall, the case-weight approach displayed strength with respect to empirical SE. To evaluate potential under- or over-estimation of the estimated SE, the relative magnitude of the mean estimated SE to empirical SE was evaluated (**Figure 4**). As a result, the case-weight and

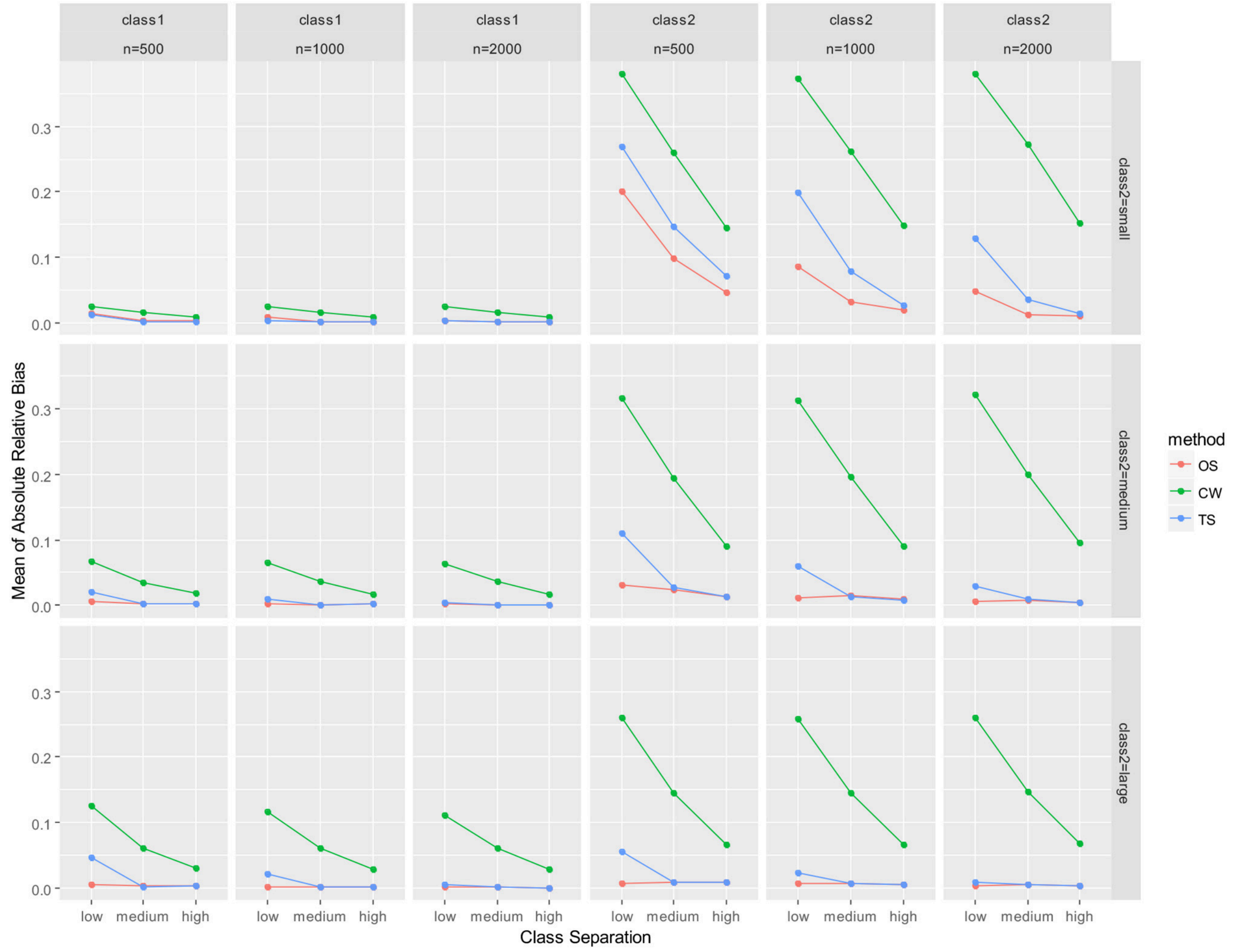


FIGURE 2 | Averaged absolute relative bias for auxiliary model parameters. OS, one-step approach; CW, case-weight approach, and TS, three-step approach.

three-step approaches displayed substantial underestimation of the SE for both classes particularly when class separation was low. For class-1 parameters, underestimation for the two approaches became small when class separation was medium or high. However, for class-2 parameters, underestimation for the case-weight approach did not diminished under small class-2 proportion conditions. Another notable result for the underestimation of the estimated SE was that the one-step approach displayed substantial underestimation for both class-1 and-2 parameters under the most demanding condition ($n = 500$, small class-2 proportion, and low separation) compared to the case-weight and three-step approaches.

As empirical SEs were evaluated for each parameter for $n = 500$ conditions (again, details are not presented here), they were notably high for the mean and variance of the intercept for class 2 by the one-step and three-step approaches when the class separation was small and the class-2 proportion was small. As the class-2 proportion became larger, empirical SE values improved for the one-step and three-step approaches, however, empirical SE values were still constantly lower by the case-weight approach.

With respect to RMSE (**Figure 5**), the performance of the one-step and three-step approaches were nearly identical and slightly better than the case-weight approach under medium/high class separation and medium/large class-2 proportion conditions. When class-2 proportion was small, the one-step approach performed slightly better than the three-step approach for larger sample sizes ($n = 1{,}000$ and 2,000). The case-weight approach performed better than the other two approaches in limited conditions. First, under small class-2 proportion conditions with $n = 500$, the case-weight approach performed constantly better than the other two approaches. Also, the case-weight approach performed better than the other two approaches under small class-2 proportion and low class separation condition with $n = 1{,}000$.

When RMSE were evaluated for each parameter for $n = 500$ conditions (again, not presented here), they were constantly

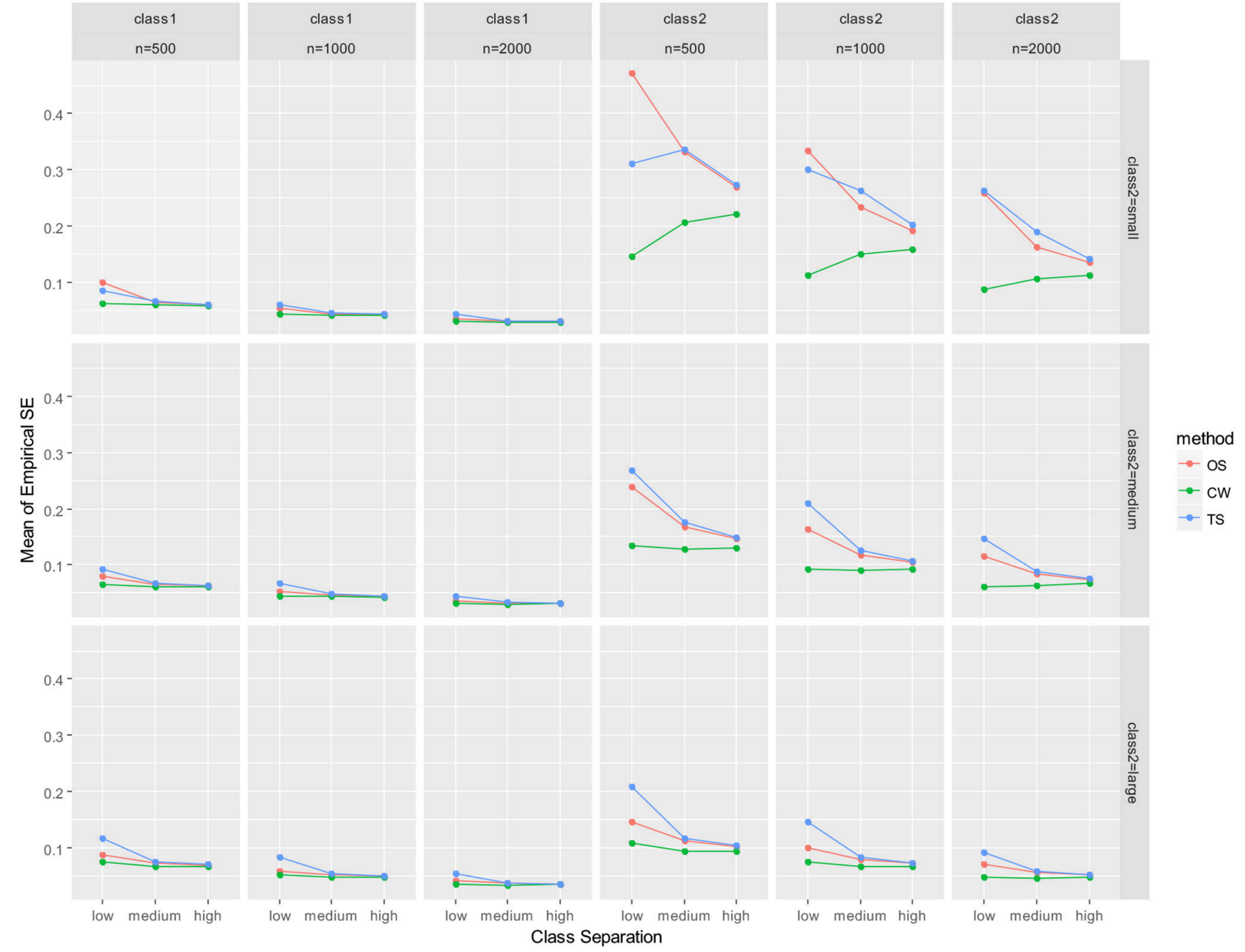


FIGURE 3 | Averaged empirical SE for auxiliary model parameters. OS, one-step approach; CW, case-weight approach, and TS, three-step approach.

low for class-1 parameters for all three approaches. For class-2 parameters, the case-weight approach constantly performed better than the other two approaches for three parameters; latent factor covariance (i.e., covariance between intercept and slope), the mean and variance of the intercept. However, the case-weight approach constantly performed worse than the other two approaches for the mean of the slope.

CONCLUSIONS

This study investigated the performance of three selected approaches for estimating two-phase mixture model, where the first phase was a two-class LCA model and the second phase was a LGM with four time points. There were some important observations in relation to the literature. First, according to Asparouhov and Muthén (2014), the loss of efficiency for the three-step approach would be minimal, compared to the one-step approach. Our results confirmed that this was the case. On the other hand, according to Asparouhov and Muthén (2014) and Vermunt (2010), parameters of the LCA model may be affected by auxiliary models, if the strength of the associations between the latent class indicators and latent classes are not sufficiently strong. This made us anticipate that parameter recovery for one-step approach would suffer in conditions with low class separations. Also, it was our hope that the case-weight approach and/or three-step approach would show better results than the one-step approach. However, it was not the case with respect to bias. One-step approach was less affected by low class separation. Also, our results displayed substantial underestimation of estimated SE for the case-weight and three-step approaches in certain conditions, which is consistent with Clark and Muthén (2009), Vermunt (2010) and Bakk et al. (2014).

PRACTICAL IMPLICATIONS

Some practically important results were demonstrated in this study. First, it was revealed that case-weight approach displayed constantly larger bias than the other two approaches. It should

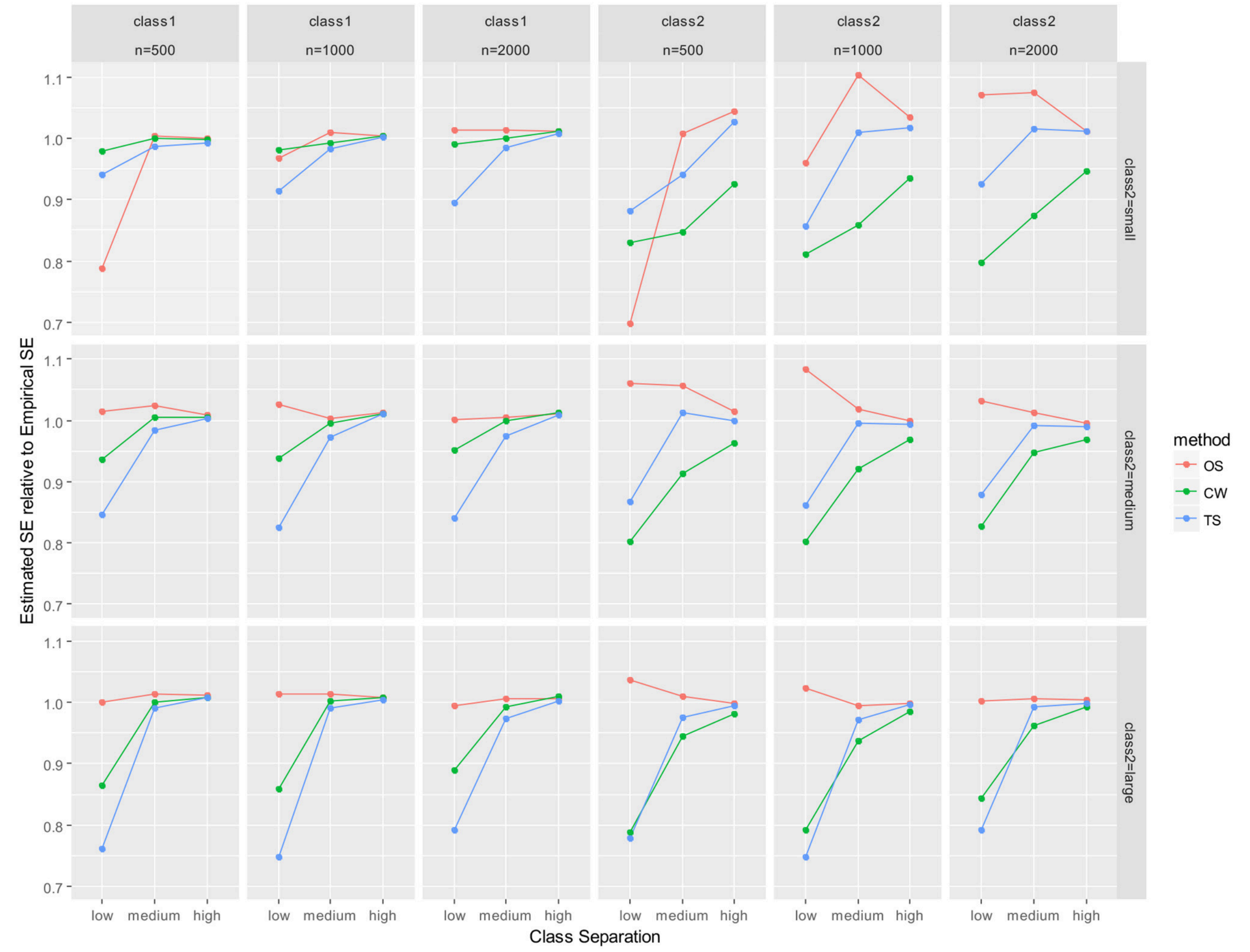


FIGURE 4 | Averaged mean estimated SE relative to mean empirical SE for auxiliary model parameters. OS, one-step approach; CW, case-weight approach, and TS, three-step approach.

be noted that this is a critical limitation of the case-weight approach. On the other hand, one-step and three-step approaches displayed much smaller bias. Their bias values were nearly identical especially when class separation was medium or high. However, their biases were high, when class-2 proportion was small, class separation was low, and the sample size was not large. Second, it was found that the case-weight approach had a strength with respect to empirical SE. However, one should be cautioned that estimated SEs were quite underestimated by the case-weight approach. Also, correct model selection rates were extremely low in such demanding conditions for all approaches, including the case-weight approach. Therefore, in practice one may not be able to take advantage of the strength of the case-weight approach with respect to SE, because there will be a lot of uncertainty in correct model selection in such demanding conditions.

Regarding successful convergence, it was found that one-step approach was very sensitive to demanding conditions. Practically, this will make one-step approach difficult to use unless the data are from ideal conditions, such as large sample size, medium to high class separation, and no presence of small class proportion. On the other hand, convergence rate was a strength of the case-weight approach under the demanding conditions. This strength makes case-weight approach allow one to explore and test more model options even in less ideal conditions. However, the case-weight approach should be used with caution in practice, because it come with substantially larger bias than the other two approaches.

Based on the results of this study, our recommendation for an application of a two-phase mixture model is as follows. First, ensure that the sample size is sufficiently large, a minimum of 500, as Asparouhov and Muthén (2014) and Vermunt (2010) have already suggested. Second, fit the latent-class measurement model part by itself to explore the number of latent classes. This makes sense because this study has demonstrated that the first-step LCA would identify a correct model better than the one-step approach. Also in this stage, it is recommended

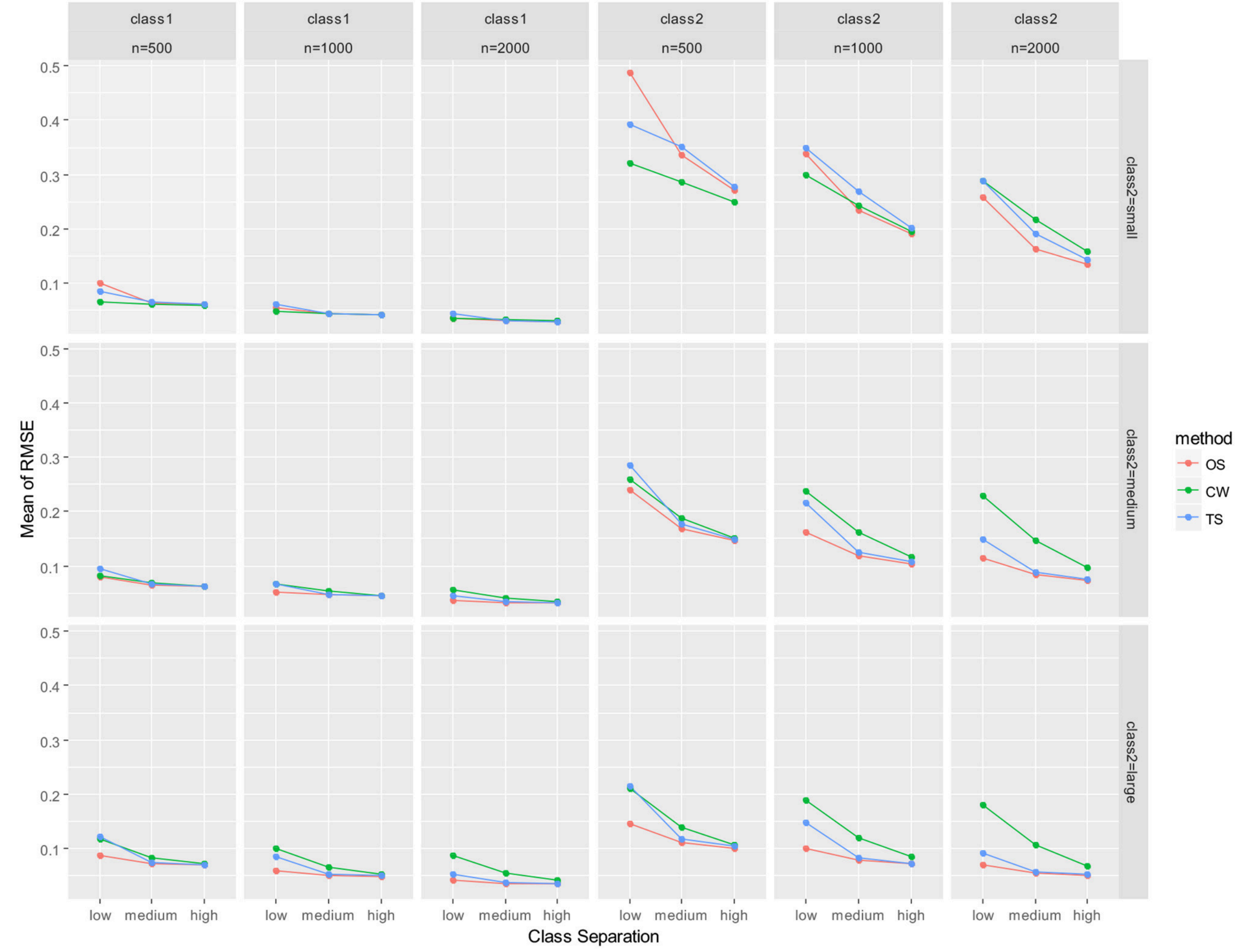


FIGURE 5 | Averaged RMSE for auxiliary model parameters. OS, one-step approach; CW, case-weight approach, and TS, three-step approach.

to ensure (a) the class separation is reasonably high, such as entropy >0.80, (b) there is no small class with <15%, to utilize the three-step approach. If these two conditions are not met, or sample size is not as large as 2,000, it is recommended to implement the one-step approach. However, if these conditions become more challenging (lower class separation and presence of smaller class), the one-step approach and the three-step approach may not converge. If so, it is when the case-weight approach is recommended to be fit. However, even if the case-weight approach converges, the results should be used with caution.

LIMITATIONS

The investigated model in this study was limited to a very specific model. As mentioned earlier in this paper, the case-weight approach and three-step approach can be applied to any kind of latent-class measurement model and any kind of auxiliary model. For example, Nese et al. (2017) employed this approach to study heterogeneity of the growth of emergent literacy knowledge by combining a two-class zero-inflated Poisson regression model as the latent-class measurement model phase, and a three-class growth mixture model as the auxiliary model phase. A future study to investigate the performance of the one-step, case-weight and three-step approaches in such a complex model is warranted.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2018.00130/full#supplementary-material

REFERENCES

Archambault, I., Janosz, M., Morizot, J., and Pagani, L. (2009). Adolescent behavioral, affective, and cognitive engagement in school: relationship to dropout. *J. School Health* 79, 408–415. doi: 10.1111/j.1746-1561.2009.00428.x

Asparouhov, T., and Muthén, B. (2014). Auxiliary variables in mixture modeling: three-step approaches using Mplus. *Struct. Equ. Model. Multidiscipl. J.* 21, 329–341. doi: 10.1080/10705511.2014.915181

Bakk, Z., Oberski, D., and Vermunt, J. (2014). Relating latent class assignments to external variables: standard errors for correct inference. *Pol. Anal.* 22, 520–540 doi: 10.1093/pan/mpu003

Bakk, Z., Tekle, F. T., and Vermunt, J. K. (2013). Estimating the association between latent class membership and external variables using bias-adjusted three-step approaches. *Sociol. Methodol.* 43, 272–311. doi: 10.1177/0081175012470644

Bandeen-Roche, K., Miglioretti, D. L., Zeger, S. L., and Rathouz, P. J. (1997). Latent variable regression for multiple discrete outcomes. *J. Am. Stat. Assoc.* 92, 1375–1386. doi: 10.1080/01621459.1997.10473658

Bolck, A., Croon, M., and Hagenaars, J. (2004). Estimating latent structure models with categorical variables: one-step versus three-step estimators. *Pol. Anal.* 12, 3–27. doi: 10.1093/pan/mph001

Cheng, Z. (2012). *The Relation between Uncertainty in Latent Class Membership and Outcomes in a Latent Class Signal Detection Model*, Doctoral dissertation, Columbia University, New York, NY. doi: 10.7916/D8ZP4D6S

Clark, S. L., and Muthén, B. (2009). *Relating Latent Class Analysis Results to Variables Not Included in the Analysis*. Avaliable online at: http://www.statmodel.com/download/relatinglca.pdf

Clogg, C. C. (1995). "Latent class models: recent developments and prospects for the future," in *Handbook of Statistical Modeling for the Social and Behavioral Sciences,* eds G. Arminger, C. C. Clogg, and M. E. Sobel (New York, NY: Plenum), 311–352.

Dayton, C. M., and Macready, G. B. (1998). Concomitant variable latent class analysis. *J. Am. Stat. Assoc.* 83, 173–178.

Formann, A. K. (1992). Linear logistic latent class analysis for polytomous data. *J. Am. Stat. Assoc.* 87, 476–486. doi: 10.1080/01621459.1992.10475229

Hagenaars, J. A. (1993). *Loglinear Models with Latent Variables*. London: Sage.

Hardigan, P. C. (2009). An application of latent class analysis in the measurement of falling among a community elderly population. *Open Geriatr. Med. J.* 2, 12–17. doi: 10.2174/1874827900902010012

Heijden, P., Dessens, J., and Bockenholt, U. (1996). Estimating the concomitant variable latent-class model with the EM algorithm. *J. Educ. Behav. Stat.* 31, 215–229. doi: 10.3102/10769986021003215

Hibbard, J. H., Mahoney, E. R., Stock, R., and Tusler, M. (2007). Do increases in patient activation result in improved self-management behaviors? *Health Serv. Res.* 42, 1443–1463. doi: 10.1111/j.1475-6773.2006.00669.x

Hirano, K., and Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: an application to data on right heart catheterization. *Health Serv. Outcomes Res. Methodol.* 2, 259–278. doi: 10.1023/A:1020371312283

Kim, M., Vermunt, J., Bakk, Z., Jaki, T., and Van Horn, M. L. (2016). Modeling predictors of latent classes in regression mixture models. *Struct. Equ. Model. Multidiscipl. J.* 23, 601–614. doi: 10.1080/10705511.2016.1158655

Lazarsfeld, P. F., and Henry, N. W. (1968). *Latent Structure Analysis*. Boston, MA: Houghton Mifflin.

Loken, E. (2004). Using latent class analysis to model temperament types. *Multivariate Behav. Res.* 39, 625–652. doi: 10.1207/s15327906mbr3904_3

Muthén, B. (2001). "Latent variable mixture modeling," in *New Developments and Techniques in Structural Equation Modeling,* eds G. A. Marcoulides and R. E. Schumacker (Hillsdale, NJ: Erlbaum), 1–33.

Muthén, B., and Muthén, L. (2000). Integrating person-centered and variable-centered analysis: growth mixture modeling with latent trajectory classes. *Alcohol. Clin. Exp. Res.* 24, 882–891. doi: 10.1111/j.1530-0277.2000.tb02070.x

Muthén, L., and Muthén, B. (1998–2012). *Mplus User's Guide. 7th Edn*. Los Angeles, CA: Muthén&Muthén.

Nese, J. F., Kamata, A., and Tindal, J. (2017). A two-step sampling weight approach to growth mixture modeling for emergent and developing skills with distributional changes over time. *J. School Psychol.* 61, 55–74 doi: 10.1016/j.jsp.2016.12.001

Nylund-Gibson, K., Grimm, R., Quirk, M., and Furlong, M. (2014). A latent transition mixture model using the three-step specification. *Struct. Equ. Model. Multidiscipl. J.* 21, 439–454. doi: 10.1080/10705511.2014.915375

R Core Team, S. (2016). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available online at: https://www.R-project.org/

Reinke, W. M., Herman, K. C., Petras, H., and Ialongo, N. S. (2008). Empirically derived subtypes of child academic and behavior problems: co-occurrence and distal outcomes. *J. Abnorm. Child Psychol.* 36, 759–770. doi: 10.1007/s10802-007-9208-2

Robins, J., and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *J. Am. Stat. Assoc.* 90, 122–129. doi: 10.1080/01621459.1995.10476494

Roeder, K., Lynch, K. G., and Nagin, D. S. (1999). Modeling uncertainty in latent class membership: a case study in criminology. *J. Am. Stat. Assoc.* 94, 766–776. doi: 10.1080/01621459.1999.10474179

Vermunt, J. K. (2010). Latent class modeling with covariates: two improved three-step approaches. *Pol. Anal.* 18, 450–469. doi: 10.1093/pan/mpq025

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 05 September 2017
doi: 10.3389/fpsyg.2017.01499

Class Enumeration and Parameter Recovery of Growth Mixture Modeling and Second-Order Growth Mixture Modeling in the Presence of Measurement Noninvariance between Latent Classes

Eun Sook Kim * *and Yan Wang*

Department of Educational and Psychological Studies, University of South Florida, Tampa, FL, United States

OPEN ACCESS

Edited by:
Suzanne Jak,
University of Amsterdam, Netherlands

Reviewed by:
Kim De Roover,
Tilburg University, Netherlands
Dylan Molenaar,
University of Amsterdam, Netherlands

***Correspondence:**
Eun Sook Kim
ekim3@usf.edu

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *22 December 2016*
Accepted: *18 August 2017*
Published: *05 September 2017*

Citation:
Kim ES and Wang Y (2017) Class Enumeration and Parameter Recovery of Growth Mixture Modeling and Second-Order Growth Mixture Modeling in the Presence of Measurement Noninvariance between Latent Classes. Front. Psychol. 8:1499. doi: 10.3389/fpsyg.2017.01499

Population heterogeneity in growth trajectories can be detected with growth mixture modeling (GMM). It is common that researchers compute composite scores of repeated measures and use them as multiple indicators of growth factors (baseline performance and growth) assuming measurement invariance between latent classes. Considering that the assumption of measurement invariance does not always hold, we investigate the impact of measurement noninvariance on class enumeration and parameter recovery in GMM through a Monte Carlo simulation study (Study 1). In Study 2, we examine the class enumeration and parameter recovery of the second-order growth mixture modeling (SOGMM) that incorporates measurement models at the first order level. Thus, SOGMM estimates growth trajectory parameters with reliable sources of variance, that is, common factor variance of repeated measures and allows heterogeneity in measurement parameters between latent classes. The class enumeration rates are examined with information criteria such as AIC, BIC, sample-size adjusted BIC, and hierarchical BIC under various simulation conditions. The results of Study 1 showed that the parameter estimates of baseline performance and growth factor means were biased to the degree of measurement noninvariance even when the correct number of latent classes was extracted. In Study 2, the class enumeration accuracy of SOGMM depended on information criteria, class separation, and sample size. The estimates of baseline performance and growth factor mean differences between classes were generally unbiased but the size of measurement noninvariance was underestimated. Overall, SOGMM is advantageous in that it yields unbiased estimates of growth trajectory parameters and more accurate class enumeration compared to GMM by incorporating measurement models.

Keywords: growth mixture modeling, second-order growth mixture modeling, measurement invariance, latent class, class enumeration

INTRODUCTION

In educational and psychological research the change or growth in temporal outcomes (e.g., alcohol use, depression, antisocial behavior, reading skills over time) is one of the major research questions (e.g., Muthén et al., 2000; Li et al., 2001; Miner and Clarke-Stewart, 2008). Given that the growth over time is likely variant across units of analysis (e.g., children), researchers are often interested in clustering in terms of the pattern or trend of growth. To investigate potential unobserved groups or latent classes in growth trajectories growth mixture modeling (GMM) is often used. For example, using GMM Baams et al. (2014) found resilients, undercontrollers, and overcontrollers in personality types; Hill et al. (2017) identified mild, increasing, elevated, and decreasing trajectories of depressive symptoms; and Oshri et al. (2017) observed declining, ascending, and stable high self-esteem.

Like many statistical methods, GMM is based on statistical assumptions. It is generally expected that the results of a statistical method are compromised to the extent to which statistical assumptions of the method are violated. One of the major assumptions of GMM is measurement invariance of longitudinal outcomes across latent classes that emerge from the data (Grimm and Ram, 2009). However, it is not known how the violation of the measurement invariance assumption impacts the performance of GMM. Thus, this study investigated the behaviors of GMM under the violation of measurement invariance across latent classes. Furthermore, we proposed the second-order growth mixture modeling (SOGMM) that allows modeling and testing measurement invariance explicitly across latent classes in the growth mixture analysis.

In the following section we first introduced latent growth modeling (LGM) that is a basic building block of second-order LGM and, next, discussed the advantages of second-order LGM addressing measurement invariance issues between observed groups in LGM. Then, we shift the focus to GMM for unobserved groups in growth trajectories and its extension to second-order GMM raising the issues of measurement noninvariance across latent classes.

Latent Growth Modeling and Second-Order Latent Growth Modeling

When researchers are interested in changes of individuals over time (e.g., changes in social role functioning over time in developmental psychology), LGM is often employed. LGM is appropriate to address research questions about the (a) average baseline performance, (b) average growth trajectories, (c) variability in baseline performance, and (d) variability in growth trajectories across individuals. That is, in addition to estimating the mean level of initial performance and growth, it allows those growth parameters to randomly vary across individuals (i.e., random effects). For example, in a study investigating the development of depressive symptoms of 7th graders over 3 years, the average depressive symptoms at grade 7 and the average growth rate of depressive symptoms over 3 years can be estimated with LGM. In addition, psychologists will be informed of how much variability exists among adolescents in terms of their initial depressive symptoms and growth rates.

In LGM, researchers can also incorporate covariates to explain the variability in the baseline scores and growth rates of depressive symptoms. For example, when gender difference is expected in the development of depressive symptoms among adolescents, this effect can be modeled and tested in LGM as shown in **Figure 1A**: gender differences in baseline depressive symptoms and growth trajectories (paths a and b, respectively). In estimating these gender differences, LGM assumes measurement invariance of depressive symptoms between boys and girls. In other words, it is assumed that boys and girls respond to the items of a depressive symptoms checklist in the same manner.

However, this assumption of measurement invariance between boys and girls can be violated, which is illustrated in **Figure 1B**. In this figure, gender differences are present not only in the initial performance and growth rates of adolescents but also in their responses to an item that measures depressive symptoms (denoted by path e). When measurement invariance between boys and girls is violated, it is well-known that the mean comparison between them is not legitimate. Generally, scalar measurement invariance (i.e., invariance of factor structure, factor loadings, and intercepts of a measurement model) is required for meaningful mean comparisons between groups (Millsap and Kwok, 2004). Specifically in the context of LGM, Kim and Willson (2014a) investigated the impact of measurement noninvariance between groups on the performance of LGM and demonstrated that intercept noninvariance was directly associated with bias and Type I error inflation on the group effect on baseline performance (path a in **Figure 1A**) whereas factor loading noninvariance was associated with bias and Type I error inflation on the growth rate (path b in **Figure 1A**). To explicitly test measurement invariance in LGM, they recommended the second-order LGM (SOLGM).

As shown in **Figure 1B**, SOLGM includes measurement models of longitudinal outcome variables as the first-order part (McArdle, 1988; Meredith and Tisak, 1990). In LGM, the temporal outcomes are observed variables measured repeatedly over a period of time (squares denoted by T_1–T_4 in **Figure 1A**). When multiple items are used to measure the outcome (e.g., depressive symptoms), it is a common practice to use composite scores of the items (Leite, 2007). When composite scores are created, all items in a measure are equally weighted regardless of their relation to the latent factor measured. On the other hand, in SOLGM the temporal outcomes are latent factors that are measured by multiple items (circles denoted by T_1–T_4 in **Figure 1B**). Thus, the relations of items to the factors are explicitly modeled with different weights (i.e., factor loadings). Because unique factor variance or error variance is taken out, growth parameters are estimated with reliable sources of variance (common factor variance; McArdle, 1988; Grimm and Ram, 2009). In addition, measurement invariance (both longitudinal invariance and group invariance) can be examined with SOLGM (Kim and Willson, 2014b).

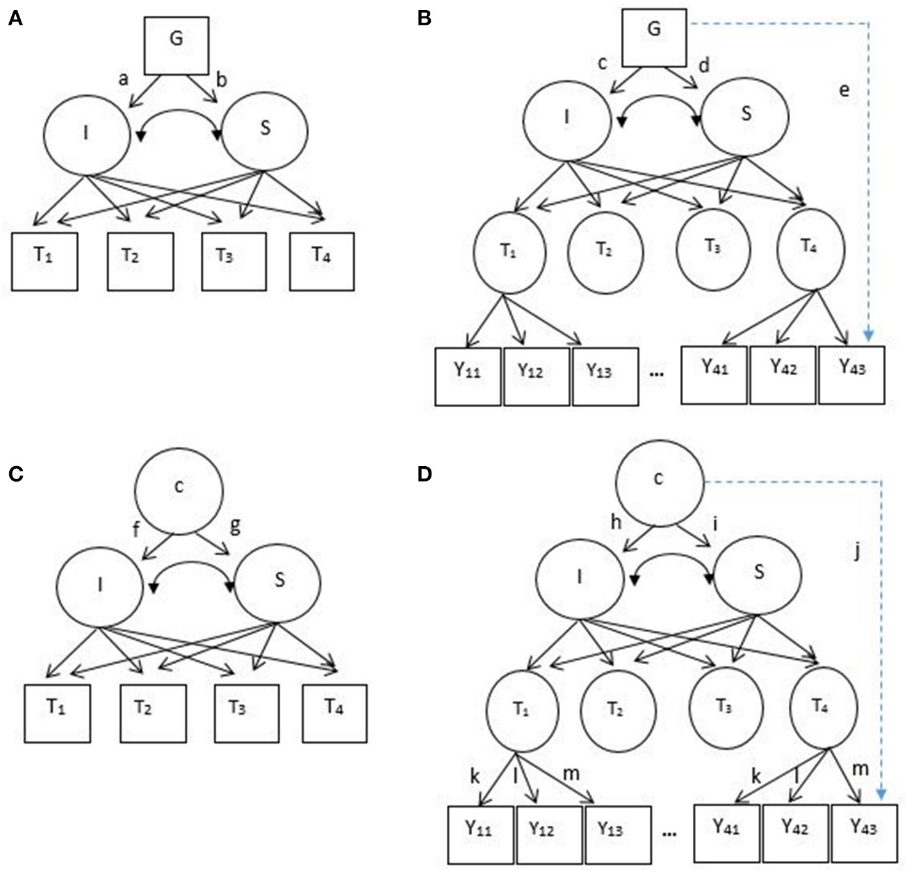


FIGURE 1 | (A) Latent growth model, LGM **(B)** second-order latent growth model, SOLGM **(C)** growth mixture model, GMM **(D)** second-order growth mixture model, SOGMM. I = continuous latent intercept, S, continuous latent slope; c, unobserved categorical variable or latent classes; G, observed covariate (e.g., gender). T_1–T_4 are observed longitudinal outcome variables (squares) in LGM and GMM, but latent factors (circles) in SOLGM and SOGMM. Y_{11}–Y_{43} are observed items of latent factors, T_1–T_4. Note that Y_{21}–Y_{33} are not shown due to a limited space. Paths a–d represent covariate effects on the intercept and slope factors (or group-specific effects if a covariate is categorical). Paths f–i represent class-specific effects on the intercept and slope factors. Path e (a dotted line) represent a covariate effect on an item (measurement noninvariance in terms of a covariate). Path j (a dotted line) represent a class-specific effect on an item (measurement noninvariance between latent classes).

Growth Mixture Modeling

Although, LGM is very useful providing information of the average initial performance and growth trajectory, it is generally assumed that all individuals are from a single population and thus the same growth pattern is applied to all individuals (Muthén, 2004; Frankfurt et al., 2016). However, it is often observed in social sciences that individuals change over time and those changes are not homogeneous across individuals. For example, the development patterns of depressive symptoms could be different among adolescents. Not to mention potential heterogeneity in baseline depressive symptoms, some may exhibit stably low or high symptoms over time, some may show steadily increasing trend, and others may experience exponential change. GMM allows researchers and practitioners to investigate the heterogeneity of growth patterns across individuals by combining the latent class approach or mixture modeling to LGM (Grimm and Ram, 2009; Frankfurt et al., 2016). Latent classes are unobserved groups that emerge from the data depending on the patterns of growth in GMM. Subgroups with their own unique growth parameters are identified as illustrated in **Figure 1C** (latent classes *c* represented by a circle as an unobserved categorical variable and their specific means of intercept and slope factors denoted by paths f and g). For example, Cabrera et al. (2016) identified four distinctive trajectories of post-combat aggression among American combat team soldiers returned from an Iraq deployment: low-stable, delayed, recovery, and chronic. They expected that their study findings could help targeted intervention of combat-related posttraumatic stress disorder through improved identification of at-risk subgroups. In addition to identifying subpopulations of heterogeneous growth curves, GMM is used to approximate non-normal distributions (McLachlan and Peel, 2000; Lubke and Neale, 2006). Normal distribution is typically assumed within subpopulation (Muthén, 2004) and the distribution of observed variables is the mixture distribution of subpopulations (Lubke and Neale, 2006). As in LGM, researchers can incorporate covariates in GMM although not demonstrated in the figure. For interested readers, refer to Muthén (2004) about the extension of GMM with covariates and distal outcomes.

When subpopulations are identified in GMM, it is assumed that measurement invariance of longitudinal outcome variables holds across identified subpopulations. For example, soldiers showing low-stable post-combat aggression and soldiers showing chronic post-combat aggression (Cabrera et al., 2016) are assumed to respond to the items of the aggression scale in the same way. However, this assumption can be violated as illustrated in **Figure 1D**. The figure shows that there is heterogeneity across latent classes not only in baseline performance and growth rates (paths h and i, respectively) but also in their responses to an item (path j). As discussed in the LGM section above, it is well-known that scalar invariance between groups is a prerequisite to a meaningful group mean comparison. Similarly, a comparison between latent classes in terms of means of intercept and slope factors (or initial performance and growth) is expected to be meaningfully interpretable when measurement invariance holds. Based on the findings of Kim and Willson (2014a) with LGM, when measurement noninvariance across latent classes is present but invariance is assumed in GMM, it is possible that spurious heterogeneity in growth parameters occurs, which can result in the detection of a spurious latent class. It is also reported that assumption violations could lead to the misidentification of an extra latent class (Bauer and Curran, 2003). Particularly, specifying a more restrictive model than a true population model within class could result in overestimation of the number of classes (Lubke and Neale, 2008; Vermunt, 2011). For example, four latent classes may be identified when there are three distinctive classes in the population. However, the impact of measurement noninvariance across latent classes on the performance of GMM has not been systematically studied yet.

Second-Order Growth Mixture Modeling

In GMM, longitudinal outcome variables are observed variables. Many applied studies using GMM employed mean or sum composite scores of multiple items of a scale (e.g., mean of eight items of 3-point peer victimization scale, Brendgen et al., 2016; sum of five items of 4-point positive religious coping scale and sum of five items of 4-point negative religious coping scale, Hayward and Krause, 2016; sum of five items of 5-point self-esteem scale, Oshri et al., 2017; mean of 16 items of 5-point depressive symptoms, Wang et al., 2015). On the other hand, the second-order GMM (SOGMM) directly models the relation of multiple items to the factor that is repeatedly measured for changes as illustrated in **Figure 1D**. Grimm and Ram (2009) described SOGMM by decomposing the model into four components: (1) a longitudinal measurement model or longitudinal common factor model, (2) measurement invariance constraints, (3) a latent growth model, and (4) a mixture model. On top of GMM (components 3 and 4) which is the second-order part, SOGMM includes a measurement model at each time point as the first-order part (components 1 and 2). Thus, the benefits of SOLGM over LGM we listed above will equally apply to SOGMM over GMM and we do not reiterate those benefits here. Of note is that Grimm and Ram demonstrated the application of SOGMM with multiple assessment (different reporters of a measure, that is, mother, father, and teacher reports of child externalizing behavior) not with multiple items of a measure. They still used sum composite scores of multiple items of mother, father, and teacher reports, and a measurement model of each scale was not employed in their SOGMM. Following their demonstration, some applications of SOGMM included multiple assessment in the measurement model with composite scores not with multiple items of a measure (e.g., Nash et al., 2015; Lee et al., 2017). Of another note is measurement invariance constraints. When measurement invariance holds over time, invariance constraints are imposed as illustrated in Grimm and Ram and also in **Figure 1D** (denoted by k, l, and m over occasions). In their demonstration measurement invariance *across latent classes* were assumed (Grimm and Ram, 2009). That is, a path from latent classes to an item (or multiple assessment) denoted by path j was constrained at zero for all items. However, in SOGMM this path, that is, measurement invariance of the corresponding item can be explicitly tested and also freely estimated when the invariance assumption does not hold. Even though SOGMM has great flexibility in testing and modeling heterogeneity across latent classes not only in growth patterns but also in measurement models, its application is not common and very limited (only with multiple assessment not with items) up to date. Moreover, the efficacy of SOGMM in detecting heterogeneous subpopulations is unknown.

Given that research on the performance of GMM and second-order GMM in the presence of measurement noninvariance is lacking, the purpose of this study is two-fold. In Study 1, we purport to investigate the impact of measurement noninvariance on class enumeration and parameter recovery in GMM. Specifically, when population heterogeneity exists in terms of measurement parameters but the scale composite scores are used in GMM ignoring measurement noninvariance, how the violation of measurement invariance affects the class enumeration and parameter estimates is examined through a Monte Carlo simulation study. In Study 2, we examine the class enumeration accuracy and parameter recovery of SOGMM in which measurement parameters are allowed to vary across latent classes.

THEORETICAL FRAMEWORK

Latent Growth Modeling

With data from repeated measures researchers can investigate growth trajectories such as the average performance at the initial stage and the average growth rate across individuals. The common factor model in structural equation modeling (SEM) can be used to address such research interest. In the SEM framework, growth trajectories such as baseline performance and growth are modeled as latent variables. As shown in **Figure 1A**, baseline performance and growth are represented by the intercept and slope latent factors, respectively in LGM. For the sake of simplicity, we assume linear growth in this example, but the model can be easily extended to different growth curves by including additional latent factors (e.g., a quadratic factor for curvilinear growth) or freely estimating factor loadings of the slope factor. The intercept and slope ($\boldsymbol{\xi}_i$) are estimated with observed continuous outcome variables of repeated measures (T_i) for an individual i (Meredith and Tisak, 1990; Wu et al., 2009) as shown in Equation (1).

$$\boldsymbol{T}_i = \boldsymbol{\Gamma}\boldsymbol{\xi}_i + \boldsymbol{\zeta}_i \tag{1}$$

where T_i is a $m \times 1$ vector of observed variables, Γ is a $m \times r$ matrix of factor loadings, $\boldsymbol{\xi}_i$ is a vector of latent factor scores (i.e., intercept and slope values), $\boldsymbol{\zeta}_i$ is a $m \times 1$ vector of time-specific error scores for an individual *i, m* is the number of occasions, and r is the number of latent factors (2 with the intercept and slope factors). For linear growth over time, the factor loadings of the intercept and slope factors can be specified as:

$$\boldsymbol{\Gamma} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & m-1 \end{bmatrix}.$$

The factor loadings of the intercept factor are all unity and those of the slope factor increase by unity from 0 to $m - 1$ to represent linear growth over m occasions. The subscript i in the intercept and slope factors ($\boldsymbol{\xi}_i$) indicates that individuals are allowed to have different intercepts (initial performance) and slopes (linear growth rates), but the average is of focal interest. The means of two latent factors, $E(\boldsymbol{\xi}_i)$ are expressed as:

$$E(\boldsymbol{\xi}_i) = \boldsymbol{\kappa} = \begin{bmatrix} \kappa_I \\ \kappa_S \end{bmatrix} \tag{2}$$

where κ_I and κ_S represent the average baseline performance and the average growth rate across individuals, respectively. The variance covariance matrix of the latent factors is:

$$\boldsymbol{\Phi} = \begin{bmatrix} \phi_I & \\ \phi_{IS} & \phi_S \end{bmatrix}$$

where ϕ_I, ϕ_S, and ϕ_{IS} represent the variability in baseline performance and growth across individuals and the covariance between baseline performance and growth, respectively. Finally, the population mean vector μ_T and variance covariance matrix Σ_T of T_i are defined as:

$$\begin{aligned} \boldsymbol{\mu}_T &= \boldsymbol{\Gamma}\boldsymbol{\kappa} \\ \boldsymbol{\Sigma}_T &= \boldsymbol{\Gamma}\boldsymbol{\Phi}\boldsymbol{\Gamma}^{'} + \boldsymbol{\Psi} \end{aligned} \tag{3}$$

where Ψ is the variance covariance matrix of residuals ($\boldsymbol{\zeta}_i$). The residuals ($\boldsymbol{\zeta}_i$) are assumed to be multivariate normally distributed with the mean of zero and independent of each other, but the assumption of independence can be relaxed by allowing residual covariance. Of note is that in LGM applications the observed outcome variables ($\boldsymbol{T}_i$,) are typically mean or sum composite scores of a measure.

Growth Mixture Modeling

To model the differences specifically in growth trajectories across individuals, GMM incorporates latent classes in LGM. Thus, GMM includes both latent continuous variables (latent factors) and latent categorical variables (latent classes; Muthén, 2004). The latent growth model introduced in Equation (1) will be specified for each latent class as shown below.

$$(\boldsymbol{T}_i|c) = \boldsymbol{\Gamma}_c\boldsymbol{\xi}_{ic} + \boldsymbol{\zeta}_{ic} \tag{4}$$

where c denotes latent classes ($c = 1, 2, \ldots, C$). Within class, the residuals ($\zeta_i|c$) are assumed to be multivariate normally distributed with a mean vector of 0 and variance covariance matrix of Ψ_c. Accordingly, Equations (2) and (3) are rewritten as

$$\begin{aligned} \boldsymbol{\kappa_c} &= \begin{bmatrix} \kappa_{Ic} \\ \kappa_{Sc} \end{bmatrix}, \\ \boldsymbol{\mu}_{Tc} &= \boldsymbol{\Gamma_c}\boldsymbol{\kappa_c}, \\ \boldsymbol{\Sigma}_{Tc} &= \boldsymbol{\Gamma_c}\boldsymbol{\Phi}_c\boldsymbol{\Gamma}_c^{'} + \boldsymbol{\Psi}_c. \end{aligned}$$

Thus, all parameters of LGM such as the intercept and slope factor means (κ_I and κ_S) and their variances and covariance (ϕ_I, ϕ_S, and ϕ_{IS}) can be class specific in GMM. In addition, the probability that an individual belongs to each category of latent classes is estimated. Hence, the distribution of the longitudinal outcome variables is a mixture of normal distributions of latent classes as shown below:

$$f(\mathbf{T}_i) = \sum_{c=1}^{C} \pi_c \varphi_c(\boldsymbol{\mu_c}, \boldsymbol{\Sigma}_{Tc}) \tag{5}$$

where $\varphi_\mathbf{c}$ is a m-dimensional normal probability density function for class c, π_c is the proportion of participants in class c, and $\sum_{c=1}^{C} \pi_c = 1$ (Bauer, 2007).

Measurement Invariance Testing in the Second-Order Growth Mixture Model

With repeated measures, the second-order growth mixture model (SOGMM) that incorporates a measurement model at the first-order level has advantages over GMM that usually uses composite scores of repeated measures. SOGMM takes into account measurement error (residuals of items not related to a common factor) and allows researchers to evaluate psychometric qualities of a scale including measurement invariance across latent classes. Thus, SOGMM is appropriate to detecting unknown clustering due to noninvariance in measurement parameters of a scale as well as heterogeneity in growth trajectories among individuals.

As illustrated in **Figure 1D**, the first-order part of SOGMM is a measurement model at each occasion t that models the relation of observed continuous variables to latent factors (e.g., depressive symptoms items to a latent factor depressive symptoms):

$$(\boldsymbol{Y}_{it}|c) = \boldsymbol{\nu}_{tc} + \boldsymbol{\Lambda}_{tc}\boldsymbol{\eta}_{itc} + \boldsymbol{\varepsilon}_{itc} \tag{6}$$

where conditional on latent class c, Y_{it} is a $p \times 1$ vector of continuous observed variables (or items), $\boldsymbol{\nu}_{tc}$ is a $p \times 1$ vector of item intercepts, $\boldsymbol{\Lambda}_{tc}$ is a $p \times q$ matrix of item factor loadings, $\boldsymbol{\eta}_{itc}$ is a $q \times 1$ vector of latent factor scores, $\boldsymbol{\varepsilon}_{itc}$ is a $p \times 1$ vector of the corresponding item error scores for an individual i, and p and q are the number of items and the number of factors, respectively. Within class, the residuals are assumed to be multivariate normally distributed with a mean vector of 0: $(\boldsymbol{\varepsilon}_{it}|c) \sim N(0, \boldsymbol{\Theta}_{tc})$.

Because measurement models are explicit in SOGMM, measurement invariance over time can be specified. Strict measurement invariance holds over time for class c if

$$\boldsymbol{\Lambda}_{tc} = \boldsymbol{\Lambda}_c, \;\; \boldsymbol{\nu}_{tc} = \boldsymbol{\nu}_c, \;\; \boldsymbol{\Theta}_{tc} = \boldsymbol{\Theta}_c.$$

Similarly, measurement invariance across latent classes can be specified in SOGMM. If measurement invariance over time holds as shown above, strict measurement invariance across latent classes ($c = 1, 2, \ldots, C$) can be further defined as:

$$\boldsymbol{\Lambda}_c = \boldsymbol{\Lambda}, \tag{7}$$

$$\boldsymbol{\nu}_c = \boldsymbol{\nu}, \tag{8}$$

$$\boldsymbol{\Theta}_c = \boldsymbol{\Theta}. \tag{9}$$

When strict invariance holds across classes, factor variances and means are freely estimated and compared across classes (or over time).

Then, the second-order part of SOGMM is basically GMM that is shown in Equation (4). To thread the first- and second-order parts of SOGMM together, the measurement model at occasion t in Equation (6) is rewritten as a measurement model over t occasions without the t subscript: $(\mathbf{Y}_i|c) = \boldsymbol{\nu}_c + \boldsymbol{\Lambda}_c \boldsymbol{\eta}_{ic} + \boldsymbol{\varepsilon}_{ic}$. By replacing η_{ic} with $(T_i|c)$ in Equation (4), the measurement model and the growth mixture model (Equation 4) are combined as the second-order growth mixture model (**Figure 1D**):

$$(\boldsymbol{Y}_i|c) = \boldsymbol{\nu}_c + \boldsymbol{\Lambda}_c \left(\boldsymbol{\Gamma}_{\mathbf{c}} \boldsymbol{\xi}_{ic} + \boldsymbol{\zeta}_{ic}\right) + \boldsymbol{\varepsilon}_{ic}.$$

The mean vector μ_{Yc} and variance covariance matrix $\boldsymbol{\Sigma}_{Yc}$ of $(\boldsymbol{Y}_i|c)$ are defined as

$$\begin{aligned} \boldsymbol{\mu}_{Yc} &= \boldsymbol{\nu}_c + \boldsymbol{\Lambda}_c \boldsymbol{\Gamma}_c \boldsymbol{\kappa}_c, \\ \boldsymbol{\Sigma}_{Yc} &= \boldsymbol{\Lambda}_c \left(\boldsymbol{\Gamma}_c \boldsymbol{\Phi}_c \boldsymbol{\Gamma}_c^{'} + \boldsymbol{\Psi}_c\right) \boldsymbol{\Lambda}_c^{'} + \boldsymbol{\Theta}_c. \end{aligned} \tag{10}$$

When GMM is used with composite scores of repeated measures, measurement noninvariance, if present, cannot be properly modeled. Kim and Willson (2014a) showed that when measurement noninvariance was present between groups but not correctly modeled by constructing latent growth models with composite scores, ignoring measurement noninvariance resulted in biased estimates of baseline performance and growth factor means and incorrect statistical inferences on these parameters. Specifically, they found that noninvariance in factor loadings led to a spurious mean difference in growth between groups whereas noninvariance in intercepts yielded a spurious mean difference in baseline performance. The size of measurement noninvariance ignored in LGM was directly related to the size of bias in those mean differences. In Appendix (Supplementary Material) we analytically demonstrated the impact of ignored measurement noninvariance on the estimates of growth factor means using SOGMM.

Class Enumeration in Growth Mixture Modeling

In practice of mixture modeling, a series of models with an increasing number of latent classes are specified. Then, the number of latent classes is commonly determined by identifying the best-fitting model among all specified models through model comparisons. To select the best-fitting model, different methods are introduced in the literature. Tein et al. (2013) summarized class enumeration methods into three categories: (a) using information criterion (IC) such as the Akaike Information Criterion (AIC; Akaike, 1974), Consistent AIC (CAIC; Bozdogan, 1987), Bayesian Information Criterion (BIC; Schwarz, 1978), and sample-size adjusted BIC (saBIC; Sclove, 1987), (b) conducting likelihood ratio tests (LRT) such as Lo-Mendell-Rubin LRT and bootstrap LRT, and (c) using entropy that evaluates how well the classes are separated. In this study we use information criteria for class enumeration. Among ICs, Nylund et al. (2007) recommended BIC and saBIC for class enumeration in GMM. These two ICs are also commonly used and suggested in the general mixture modeling literature (e.g., Lubke and Muthén, 2005; Tay et al., 2011). However, some authors showed the outperformance of AIC over BIC particularly when sample size was small and the class separation was poor (Lukočienė et al., 2010), but the AIC tended to overestimate the number of latent classes in other cases (Celeux and Soromenho, 1996; Nylund et al., 2007; Tein et al., 2013). In model selection with mixture modeling, the hierarchical BIC (HBIC) is also suggested (Zhao et al., 2013, 2015; Gollini and Murphy, 2014; Zhao, 2014). Zhao et al. (2015) argued that BIC tends to overpenalize model complexity in mixture modeling by using the total sample size for all estimated parameters and suggested to penalize parameters with their relevant sample size, that is, local or effective sample size that is used to estimate parameters associated with a specific class ($n\pi_c$ in the equation below). As shown in Equation (11), the HBIC equals to the BIC when $c = 1$, but is smaller than BIC when $c > 1$. Zhao and colleagues demonstrated that the HBIC outperformed the BIC especially when sample size was small. Thus, we included these four ICs in our study. These ICs are computed as:

$$\begin{aligned} \text{AIC} &= -2\text{log}L + 2^{*}k, \\ \text{BIC} &= -2\text{log}L + log(n)^{*}k, \\ \text{saBIC} &= -2\text{log}L + log[(n+2)/24]^{*}k, \\ \text{HBIC} &= -2\text{log}L + \left(k_0 + C - 1\right)\log(n) + \sum_{c=1}^{C} \log(n\pi_c) * k_c^{'} \end{aligned} \tag{11}$$

where log$L$ means log likelihood, log(n) is the natural logarithm of sample size, log($n\pi_c$) is the natural logarithm of sample size specific to a latent class c where $\pi_c \geq 0$, $c = 1, 2, \ldots, C$, and $\sum_{c=1}^{C} \pi_c = 1$, k and $k_c^{'}$ represent the number of freely estimated parameters for the total sample and for a latent class c, respectively, and k_0 is the number of free parameters common across latent classes (hence, $k = k_0 + C - 1 + \sum_{c=1}^{C} k_c^{'}$).

This study investigated how measurement noninvariance in a scale across latent classes makes impact on class enumeration and parameter estimates when GMM is used to evaluate growth over time ignoring the lack of invariance. In addition, when SOGMM is used, that is, measurement models are incorporated in GMM and measurement parameters are allowed to be heterogeneous across latent classes, the class enumeration accuracy and bias of parameter estimates in SOGMM was examined in the presence of measurement noninvariance. We hypothesize the following:

1. When baseline performance and growth are homogeneous on average, that is, latent classes are not present in terms of the intercept and slope factor means, GMM would falsely identify latent classes because the ignored measurement noninvariance would be detected as heterogeneity in these factor means.
2. When latent classes are falsely identified, a spurious mean difference in the slope factor would be observed if there is noninvariance in factor loadings; a spurious mean difference in the intercept factor would be observed if there is noninvariance in intercepts. The size of the spurious mean difference would be associated with the size of ignored measurement noninvariance.
3. SOGMM would correctly identify the number of latent classes in the presence of measurement noninvariance.
4. SOGMM would yield unbiased estimates of the difference between latent classes with respect to the intercept and slope factor means in the growth model part as well as factor loadings and intercepts in the measurement model part.

STUDY 1: GROWTH MIXTURE MODELING IN THE PRESENCE OF MEASUREMENT NONINVARIANCE BETWEEN LATENT CLASSES

Method

We conducted a Monte Carlo simulation study to investigate the impact of measurement noninvariance on the class enumeration and parameter recovery of GMM. The simulation factors included (a) location of noninvariance (factor loading/intercept), (b) degree of noninvariance (small/large), (c) difference in the intercept and slope factor means (zero/large), (d) sample size (100/200/400/1000), and (e) mixing proportion (balanced/unbalanced). Because the impact of measurement noninvariance on the performance of GMM was of focal interest in this study, the following factors were fixed as a constant for simplicity of discussions: two latent classes, four occasions, six items that load on a single factor at each occasion, and two noninvariant items, which were commonly adopted in previous simulation studies (e.g., Nylund et al., 2007; Chen et al., 2010; Kim and Willson, 2014a). In addition, measurement invariance over time was simulated. Although temporal invariance can also be violated in reality, the impact of noninvariance across latent classes could be less clear to delineate when noninvariance is present at both locations. Of note is that measurement invariance over time can be tested separately with a longitudinal common factor model and, if invariance holds, researchers can impose temporal invariance constraints on SOGMM which was demonstrated by Grimm and Ram (2009). However, measurement invariance across classes cannot be tested separately because latent classes are unobservable in advance. Of another note is that we investigated the impact of measurement noninvariance in factor loadings and intercepts between latent classes (violation of Equations 7 and 8) because scalar invariance is considered as a prerequisite to meaningful mean comparisons across groups (Millsap and Kwok, 2004; Raykov et al., 2012; Jak et al., 2014). Finally, error correlations over time were not simulated for the simplicity because this is not a major interest in this study.

Data Generation

Data were generated using the second-order growth mixture model with a measurement model at each occasion of repeated measures. The population parameters used for data generation are presented in **Figure 2**. The parameters in the first-order measurement model were majorly adopted from Kim and Willson (2014a) who conducted a similar study with observed groups using the second-order LGM. The generated values of factor loadings ($0.80 \sim 1.25$), intercepts ($-0.15 \sim 0.25$), and residual variances (0.36) were also observed in previous simulation studies of measurement invariance (Wirth, 2009; Kim et al., 2012). In the second-order latent growth model, a linear growth over four occasions was simulated. The means of baseline performance and growth factors (or intercept and slope factors) were 0 and 1, respectively. The respective variances were 0.5 and 0.1, and their covariance was 0.089 which corresponded to correlation 0.4 (Leite, 2007). The ratio of the intercept factor variance to the slope factor variance, 5:1 is considered reasonable in practice and adopted in other simulation studies (Muthén and Muthén, 2002; Depaoli, 2013; Li and Harring, 2016). The reliability estimates of 24 generated items (6 items per occasion) ranged from 0.59 to 0.91.

In the first-order measurement model, two out of six items were simulated as noninvariant across all simulation conditions. The 0.20 and 0.40 differences between classes for small and large factor loading noninvariance, respectively, and the 0.30 and 0.60 differences for small and large intercept noninvariance, respectively, were generated (e.g., Stark et al., 2006). On top of measurement noninvariance, population heterogeneity was simulated in the mean of intercept and slope factors at two levels (zero or large). When there was no difference between two classes in the intercept and slope factor means, both classes had the intercept and slope factor means of 0 and 1, respectively. When a large difference between latent classes was generated, the intercept and slope factor means of the second class were higher by 1.4 and 0.4, respectively and thus this class performs better at the baseline and also grows faster over time. The generated mean differences corresponded to Mahalanobis distance (MD) 2.0. In the mixture literature, MD 2.0 is considered as large class

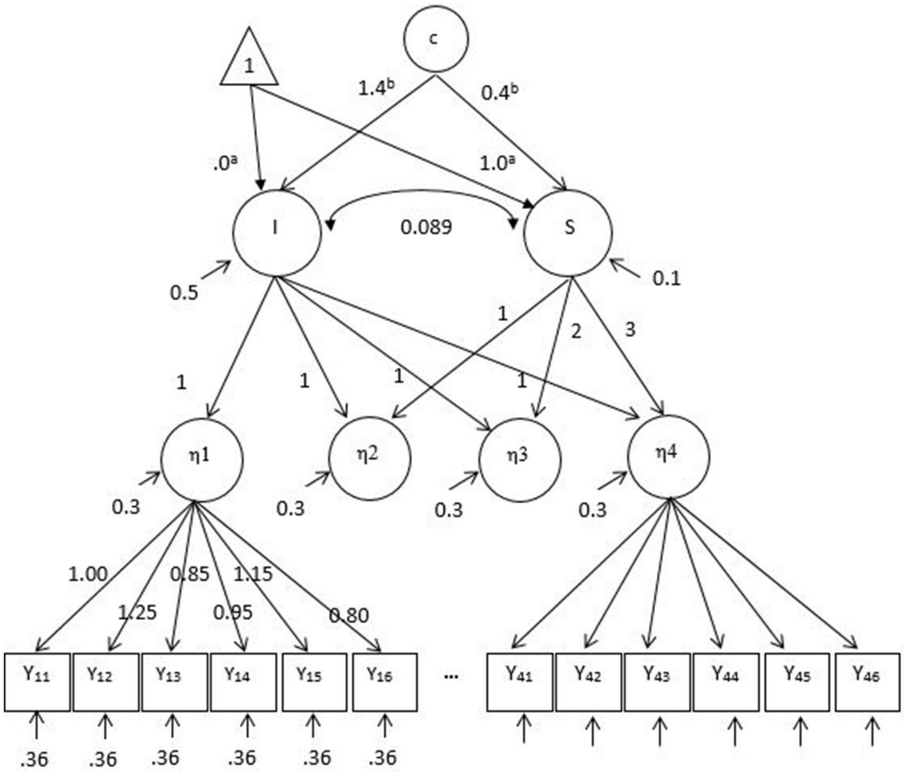


FIGURE 2 | Population parameters for data generation with the second-order growth mixture model under the factor mean difference conditions. A linear growth over time is generated. I, Latent intercept; S, latent slope; c, unobserved categorical variable or latent classes. For simplicity the measurement intercept values are not specified in this figure. Y_{11}–Y_{46} are observed items of latent factors, η1–η4. Note that Y_{21}–Y_{36} are not shown due to a limited space. The same set of factor loadings and residual variances of six items are applied over time for η1–η4. [a]The intercept and slope factor means of a latent class (i.e., reference class), respectively. [b]The mean differences between latent classes for the intercept and slope factors, respectively when the number of classes is two.

separation (e.g., Tueller and Lubke, 2010; Depaoli, 2013; Li and Harring, 2016). Of note is that class separation is one of major factors associated with the correct enumeration of latent classes (Henson et al., 2007; Tofighi and Enders, 2008; Chen et al., 2010).

The combination of two simulation factors, that is, (a) noninvariance in measurement parameters and (b) mean differences in the intercept and slope factors yielded two types of population heterogeneity. When there was no factor mean difference, measurement noninvariance was the only source of population heterogeneity that differentiated two latent classes. When there were factor mean differences, two sources of population heterogeneity, that is, measurement noninvariance and factor mean differences separated two latent classes. In the latter, the latent class with higher item factor loadings had higher factor means under the factor loading noninvariance conditions; the latent class with higher item intercepts had higher factor means under the intercept noninvariance conditions. By generating data in this way (positive pairing), we expected that the factor mean differences between latent classes would be overestimated when the invariance was assumed because the ignored noninvariance could make the factor mean of the higher class even higher (the factor mean of the lower class even lower). This was illustrated in Appendix (Supplementary Material). When two latent classes were disproportionately formed, the mixing proportion was 80 and 20%. The latent class with higher factor means and/or measurement parameters was associated with a large sample size (i.e., 80%) when two latent classes were unbalanced. For each condition, 500 replications were generated using Mplus version 7.3 (Muthén and Muthén, 2014).

Fitted Models and Simulation Outcomes

In Study 1, the fitted model was a growth mixture model in which measurement noninvariance could not be modeled although the data were generated with measurement noninvariance. The mean composite scores were used as observed indicators of the growth mixture model ignoring noninvariance of items. It should be noted that equal weights were applied for all items when composite scores were created although factor loadings were different across items in the population. A linear growth was modeled with factor loadings of the intercept factor all fixed at 1 and those of the slope factor specified at 0, 1, 2, and 3 for four occasions. Because there were two latent classes in the population, models with one, two, and three latent classes were evaluated and a best fitting model was selected on the basis of the selected fit index (i.e., AIC, BIC, saBIC, and HBIC). Note that a latent class with a cell proportion <0.05 was ruled out because the number of observations in a class (e.g., less than five observations with $N = 100$) was too small (e.g., Feldman et al., 2009). For example, even though the ICs supported three classes, if one of them constituted <5% of total observations, this replication was counted as two classes. The class enumeration was recorded for each replication and the enumeration rates for one-, two-, and three-class models were computed by simulation conditions and fit indexes. The enumeration rate of two classes, for example, was computed by dividing the number of replications that supported a two-class model by the total number of replications.

When no factor mean difference was simulated in the growth model (precisely speaking, the second-order part of SOGMM), one class was considered as a correct number of classes. However, we hypothesized that two classes would emerge due to ignored measurement noninvariance between two classes as observed in Kim and Willson (2014a; that is, a factor mean difference was detected when there was no factor mean difference between two groups). When two classes were generated with different factor means in the intercept and slope factors, two classes were expected to be detected correctly. However, in either scenario (one class or two classes), we hypothesized that the parameter estimates of GMM would be biased due to ignored measurement noninvariance in the measurement model. Specifically, we examined the bias in the means of the intercept and slope factors. The bias was estimated as the average difference between the estimated factor mean and the generated population factor mean across replications. If the population parameter was not zero, we also estimated relative bias which is the ratio of the estimated bias to the population parameter. Relative bias >0.05 is typically considered substantial in the simulation studies (Hoogland and Boomsma, 1998). Standardized bias was not considered because it is possibly affected by sample size (the larger sample size, the smaller standardized bias holding raw bias constant). Although, the values of raw bias are less interpretable, raw bias would suffice to show the impact of measurement noninvariance on the estimates of GMM. GMM was fitted with Mplus version 7.3 (Muthén and Muthén, 2014).

Results

Class Enumeration

The class enumeration rates of AIC, BIC, saBIC, and HBIC for the balanced conditions are presented in **Table 1**. The enumeration rates for the unbalanced conditions are similar and, thus, not presented here. The top panel of **Table 1** showed the enumeration rates of GMM when there was no heterogeneity in the intercept and slope factor means. Thus, one class was considered as a correct number of classes. However, because GMM used mean composite scores ignoring measurement noninvariance between two latent classes, we hypothesized that two classes would emerge. Unexpectedly, one class was generally selected, which might indicate that GMM was not very sensitive to the ignored measurement noninvariance. The BIC and HBIC identified one class as a best fit model almost always regardless of simulation factors. The saBIC also selected one class, but less frequently as sample size decreased (e.g., 0.97 with $N = 1{,}000$ and 0.28 with $N = 100$ when small noninvariance was simulated in factor loadings). The AIC was not affected by simulation factors much: the enumeration rates for a one-class model were around 0.60 across simulation conditions.

The bottom panel of **Table 1** showed the enumeration rates of GMM when differences in the intercept and slope factor means were simulated between two classes in addition to measurement noninvariance. The model with two latent classes was expected to be selected. However, the BIC and HBIC still selected one class more frequently showing its insensitivity to population heterogeneity when sample size was not large. Only when sample size reached 1,000 and the size of noninvariance was large, two classes were mostly identified. We could observe the impact of the unmodeled measurement noninvariance on the enumeration rates of BIC and HBIC. Specifically, the enumeration accuracy of BIC and HBIC depended on the size and location of noninvariance. When factor loadings were noninvariant, both selected two classes more frequently compared to the intercept noninvariance conditions. As the size of noninvariance increased, the correct enumeration rates also increased. This possibly implies that the ignored measurement noninvariance created larger class separation by adding its unmodeled effect to the factor mean differences in GMM. This impact of measurement noninvariance was not observed with the AIC and saBIC. In general, the correct enumeration rates of the saBIC were higher than those of the AIC, BIC, and HBIC. The HBIC outperformed the BIC, but it sometimes over-identified latent classes (i.e., three classes). For all four information criteria, it was prominent that the enumeration accuracy was associated with sample size (the larger, the more accurate). We also examined the class proportion when two classes were selected. The class proportion was generally consistent to the population proportion (that is, 50 and 50% for the balanced conditions and 20% and 80% for the unbalanced conditions).

Bias and Relative Bias

Bias and relative bias were examined with the replications in which the number of classes was correctly identified (see **Table 2**). It should be noted that the raw bias of the intercept and slope factor means is presented for no factor mean difference conditions because the intercept factor mean is zero (i.e., relative bias cannot be computed). For factor mean difference conditions the relative bias of the differences is presented (the intercept factor mean difference = 1.4; the slope factor mean difference = 0.4). Across conditions, three patterns emerged. First, only the factor means associated with the ignored noninvariance were biased whereas the factor means not associated with the ignored noninvariance were unbiased with raw and relative bias close to zero. Specifically, when noninvariance was simulated in the factor loadings of the first-order measurement model but ignored in GMM, the slope factor means showed notable bias while the intercept factor means remained unbiased; when noninvariance was simulated in the intercepts of the first-order measurement model but ignored in GMM, the intercept factor means were biased while the slope factor means were unbiased. Second, the magnitude of bias was directly related to the magnitude of ignored measurement noninvariance irrespective of sample size. This pattern was clearly observed in the raw

TABLE 1 | The class enumeration rates of growth mixture modeling for the balanced conditions.

DIF location	DIF size	Sample size	AIC			BIC		saBIC			HBIC	
			Number of latent classes									
			1	2	3	1	2[a]	1	2	3	1	2[a]
NO FACTOR MEAN DIFFERENCE												
Loading	Small	50/50	**0.56**	0.28	0.16	**0.99**	0.01	**0.28**	0.41	0.30	**0.89**	0.09[a]
		100/100	**0.65**	0.23	0.11	**0.99**	0.01	**0.71**	0.19	0.10	**0.97**	0.02[a]
		200/200	**0.68**	0.24	0.08	**1.00**	0.00	**0.91**	0.08	0.02	**0.99**	0.01
		500/500	**0.72**	0.23	0.05	**1.00**	–	**0.97**	0.03	0.00	**0.99**	0.01
	Large	50/50	**0.55**	0.29	0.16	**0.99**	0.01	**0.28**	0.42	0.31	**0.90**	0.08[a]
		100/100	**0.64**	0.27	0.09	**1.00**	0.00	**0.69**	0.23	0.08	**0.96**	0.03[a]
		200/200	**0.67**	0.25	0.08	**1.00**	–	**0.89**	0.09	0.02	**0.99**	0.01
		500/500	**0.68**	0.27	0.05	**1.00**	0.00	**0.95**	0.05	0.00	**0.99**	0.01
Intercept	Small	50/50	**0.57**	0.26	0.17	**0.99**	0.01	**0.30**	0.39	0.31	**0.89**	0.09[a]
		100/100	**0.64**	0.25	0.11	**0.99**	0.01	**0.69**	0.23	0.08	**0.97**	0.02[a]
		200/200	**0.69**	0.25	0.06	**1.00**	0.00	**0.90**	0.08	0.02	**0.99**	0.01
		500/500	**0.74**	0.22	0.04	**1.00**	–	**0.97**	0.02	0.00	**1.00**	0.00
	Large	50/50	**0.59**	0.24	0.17	**0.99**	0.01	**0.31**	0.37	0.32	**0.90**	0.09[a]
		100/100	**0.63**	0.26	0.11	**0.99**	0.01	**0.68**	0.23	0.09	**0.97**	0.03[a]
		200/200	**0.68**	0.26	0.06	**1.00**	0.00	**0.90**	0.08	0.02	**0.99**	0.01
		500/500	**0.75**	0.20	0.05	**1.00**	–	**0.97**	0.03	–	**1.00**	0.00
FACTOR MEAN DIFFERENCE												
Loading	Small	50/50	0.36	**0.40**	0.24	0.94	**0.06**[a]	0.17	**0.44**	0.39	0.80	**0.16**[a]
		100/100	0.19	**0.63**	0.18	0.91	**0.09**	0.22	**0.63**	0.16	0.80	**0.18**[a]
		200/200	0.08	**0.75**	0.17	0.78	**0.22**	0.21	**0.72**	0.07	0.63	**0.37**[a]
		500/500	0.00	**0.80**	0.20	0.21	**0.79**	0.02	**0.94**	0.04	0.12	**0.87**[a]
	Large	50/50	0.25	**0.48**	0.27	0.87	**0.12**[a]	0.09	**0.48**	0.43	0.70	**0.24**[a]
		100/100	0.10	**0.65**	0.25	0.79	**0.21**[a]	0.13	**0.65**	0.22	0.60	**0.37**[a]
		200/200	0.01	**0.67**	0.32	0.42	**0.57**[a]	0.03	**0.79**	0.18	0.29	**0.68**[a]
		500/500	–	**0.52**	0.48	0.00	**0.98**[a]	–	**0.78**	0.22	0.00	**0.90**[a]
Intercept	Small	50/50	0.40	**0.40**	0.20	0.94	**0.06**	0.20	**0.45**	0.35	0.82	**0.16**[a]
		100/100	0.31	**0.55**	0.13	0.93	**0.07**	0.35	**0.53**	0.11	0.84	**0.16**
		200/200	0.13	**0.71**	0.17	0.83	**0.17**	0.28	**0.67**	0.05	0.73	**0.27**
		500/500	0.00	**0.90**	0.09	0.35	**0.65**	0.04	**0.94**	0.02	0.23	**0.77**
	Large	50/50	0.32	**0.46**	0.22	0.92	**0.08**[a]	0.12	**0.49**	0.39	0.78	**0.20**[a]
		100/100	0.23	**0.58**	0.19	0.88	**0.12**	0.27	**0.57**	0.15	0.78	**0.22**[a]
		200/200	0.06	**0.80**	0.15	0.74	**0.26**	0.17	**0.78**	0.05	0.61	**0.39**[a]
		500/500	–	**0.87**	0.13	0.18	**0.82**	0.01	**0.96**	0.03	0.10	**0.90**[a]

The hypothesized correct enumeration rates are in bold. DIF, Differential item functioning or measurement noninvariance; AIC, Akaike information criterion; BIC, Bayesian information criterion; saBIC, sample-size adjusted BIC; HBIC, hierarchical BIC. Due to rounding 0.00 means one or two replications out of 500.

[a] *We compared one-, two-, and three-class models, and the three-class model was selected with a small proportion.*

bias under no factor mean difference conditions. When the magnitude of noninvariance was doubled, the magnitude of bias in factor means was also doubled. For example, for the balanced conditions with ignored intercept noninvariance, the raw bias in the intercept factor means was about 0.05 when small noninvariance was ignored and about 0.10 when large noninvariance was ignored. Third, the direction of bias also reflected the direction of ignored measurement noninvariance. Under the no factor mean difference conditions, two factor loadings were simulated lower in one class, which probably led to negative bias in the slope factor means whereas two intercepts were simulated higher in one class, which conceivably resulted in positive bias in the intercept factor means. Under the factor mean difference conditions, noninvariance, regardless of its location, was simulated in favor of the class with higher factor means (positive pairing; the latent class with higher item factor loadings had higher factor means under the factor loading noninvariance conditions; the latent class with higher item intercepts had higher factor means under the intercept noninvariance conditions). As hypothesized and illustrated in Appendix (Supplementary Material), the corresponding factor mean differences were mostly positively biased because the factor mean of the class with a higher factor mean tended to be overestimated and that of the class with a lower factor mean tended to be underestimated.

TABLE 2 | The bias and relative bias of the intercept and slope factor means in growth mixture modeling.

DIF location	DIF size	Sample size	No difference (Raw bias)		Difference (Relative bias)	
			Intercept	Slope	Intercept *d*	Slope *d*
Loading	Small	50/50	0.006	−0.034	−0.003	−0.210
		100/100	0.004	−0.034	0.002	0.035
		200/200	0.005	−0.033	−0.003	0.125
		500/500	0.002	−0.033	0.009	0.165
	Large	50/50	0.006	−0.067	0.001	0.182
		100/100	0.004	−0.067	0.010	0.317
		200/200	0.004	−0.066	0.019	0.345
		500/500	0.002	−0.066	0.017	0.360
Intercept	Small	50/50	0.057	0.000	0.054	−0.510
		100/100	0.055	−0.001	0.083	−0.155
		200/200	0.055	0.001	0.089	−0.053
		500/500	0.051	0.000	0.071	0.001
	Large	50/50	0.107	0.000	0.110	−0.108
		100/100	0.105	−0.001	0.165	−0.028
		200/200	0.105	0.001	0.151	−0.020
		500/500	0.101	0.000	0.145	−0.008
Loading	Small	80/20	0.006	−0.014	0.007	−0.245
		160/40	0.004	−0.014	0.017	0.006
		320/80	0.005	−0.013	−0.009	0.114
		800/200	0.002	−0.013	−0.035	0.113
	Large	80/20	0.006	−0.027	−0.058	−0.080
		160/40	0.004	−0.027	−0.015	0.138
		320/80	0.004	−0.026	−0.044	0.237
		800/200	0.002	−0.026	−0.065	0.238
Intercept	Small	80/20	0.087	−0.001	0.081	−0.270
		160/40	0.085	−0.001	0.118	−0.043
		320/80	0.085	0.001	0.094	0.015
		800/200	0.081	0.000	0.072	0.000
	Large	80/20	0.167	−0.001	0.159	−0.220
		160/40	0.165	−0.001	0.184	−0.023
		320/80	0.165	0.001	0.155	−0.003
		800/200	0.161	0.000	0.142	0.004

DIF, Differential item functioning or measurement non-invariance; No difference, no factor mean difference; Difference, factor mean difference; Intercept, intercept factor mean; Slope, slope factor mean; Intercept d, intercept factor mean difference; Slope d, slope factor mean difference.

Compared to the balanced conditions, the parameter estimates in the unbalanced conditions were more biased when intercepts were not invariant, but less biased when factor loadings were not invariant.

STUDY 2: SECOND-ORDER GROWTH MIXTURE MODELING

Method

Fitted Models and Simulation Outcomes

The data generated in Study 1 were fitted to the second-order growth mixture models that allow heterogeneity in the measurement parameters at the first-order measurement model. The second-order part was specified identical to the GMM in Study 1. Instead of using observed mean composite scores for the indictors of the intercept and slope factors, a latent factor on which six items loaded was included at each occasion as shown in **Figure 2**. As in Study 1, we fitted one-, two-, and three-class models and decided the number of classes based on the fit criteria (lowest information criterion) applying the minimum class proportion 0.05 rule. Because two latent classes were generated in the population and SOGMM was a correctly specified model, two classes were expected to be selected across all simulation conditions. In addition, bias or relative bias in parameter estimates was evaluated. The parameters of interest included the differences between classes in the intercept and slope factor means and the size of noninvariance in the factor loadings and intercepts of two noninvariant items. The size of noninvariance was averaged across two items. It was hypothesized that SOGMM would yielded unbiased estimates of these parameters. Mplus version 7.3 (Muthén and Muthén, 2014) was used for SOGMM.

Results

Class Enumeration

The class enumeration rates of AIC, BIC, saBIC, and HBIC are presented in **Tables 3**, **4**. Because population heterogeneity was simulated between two classes either in measurement parameters only (no factor mean difference conditions in **Table 3**) or in both measurement and structural parameters (factor mean difference conditions in **Table 4**), we hypothesized that SOGMM would identify two latent classes correctly. However, the correct enumeration rates varied depending on the fit criteria and simulation factors. First, when there were differences in both measurement and structural parameters with substantial factor mean differences, the BIC and HBIC almost always endorsed two classes correctly. However, when measurement noninvariance was the only source of population heterogeneity (i.e., lower class separation), the correct enumeration rates of BIC and HBIC deteriorated notably as sample size decreased and the noninvariance size was small. For example, the BIC was totally insensitive to the small noninvariance in the intercepts and selected one class across all replications. Under these conditions, the outperformance of HBIC over BIC was observed. On the other hand, the performance of saBIC was related more with sample size but less with the magnitude of noninvariance and class separation. Thus, as sample size increased, the correct enumeration rates of saBIC reached 100% with a few exceptions in the small intercept noninvariance only conditions. Interestingly, no salient difference was observed between small and large noninvariance conditions and also between no factor mean difference and factor mean difference conditions. For the AIC, the impact of sample size was observed only when the class separation was low (i.e., small noninvariance conditions without factor mean differences). When the class separation was sufficiently large (i.e., large noninvariance conditions even without factor mean differences), the overall performance of AIC was less affected by other simulation factors showing consistent enumeration rates. Of note is that as classes were separated more

TABLE 3 | The class enumeration rates of second-order growth mixture modeling under the no factor mean difference conditions.

DIF location	DIF size	Sample size	AIC			BIC		saBIC			HBIC	
			Number of latent classes									
			1	2	3	1	2[a]	1	2	3	1	2[b]
Loading	Small	50/50	0.18	**0.77**	0.05	0.99	**0.01**	0.03	**0.80**	0.17	0.63	**0.37**
		100/100	0.03	**0.90**	0.07	0.90	**0.10**	0.03	**0.92**	0.05	0.52	**0.48**
		200/200	–	**0.93**	0.07	0.36	**0.64**	0.00	**0.99**	0.01	0.10	**0.90**
		500/500	–	**0.94**	0.06	–	**1.00**	–	**1.00**	–	–	**1.00**
	Large	50/50	–	**0.90**	0.10	–	**1.00**	–	**0.74**	0.26	–	**0.98**[b]
		100/100	–	**0.90**	0.10	–	**1.00**	–	**0.93**	0.07	–	**1.00**
		200/200	–	**0.94**	0.06	–	**1.00**	–	**0.99**	0.01	–	**1.00**
		500/500	–	**0.95**	0.05	–	**1.00**	–	**1.00**	–	–	**1.00**
Intercept	Small	50/50	0.65	**0.31**	0.04	1.00	–	0.22	**0.66**	0.12	0.84	**0.15**[b]
		100/100	0.59	**0.39**	0.02	1.00	–	0.66	**0.33**	0.01	0.92	**0.08**
		200/200	0.46	**0.51**	0.03	1.00	–	0.83	**0.16**	0.00	0.96	**0.04**
		500/500	0.08	**0.87**	0.04	1.00	**0.00**	0.70	**0.30**	–	0.95	**0.05**
	Large	50/50	0.02	**0.88**	0.09	0.80	**0.20**	0.00	**0.78**	0.22	0.29	**0.69**[b]
		100/100	–	**0.92**	0.08	0.27	**0.73**	–	**0.95**	0.05	0.06	**0.94**
		200/200	–	**0.95**	0.05	0.00	**1.00**	–	**0.99**	0.01	–	**1.00**
		500/500	–	**0.97**	0.03	–	**1.00**	–	**1.00**	–	–	**1.00**
Loading	Small	80/20	0.46	**0.51**	0.03	1.00	**0.00**	0.12	**0.75**	–	0.77	**0.23**
		160/40	0.25	**0.71**	0.04	1.00	**0.00**	0.31	**0.67**	–	0.79	**0.21**
		320/80	0.06	**0.89**	0.05	0.95	**0.05**	0.21	**0.78**	0.01	0.54	**0.46**
		800/200	0.00	**0.95**	0.05	0.32	**0.68**	0.01	**0.99**	–	0.03	**0.97**
	Large	80/20	–	**0.93**	0.07	0.08	**0.92**	–	**0.75**	0.25	0.01	**0.99**[b]
		160/40	–	**0.93**	0.07	–	**1.00**	–	**0.95**	0.05	–	**1.00**[b]
		320/80	–	**0.92**	0.08	–	**1.00**	–	**1.00**	0.00	–	**1.00**[b]
		800/200	–	**0.96**	0.04	–	**1.00**	–	**1.00**	–	–	**1.00**
Intercept	Small	80/20	0.72	**0.25**	0.03	1.00	–	0.23	**0.64**	0.13	0.85	**0.15**[b]
		160/40	0.70	**0.29**	0.01	1.00	–	0.77	**0.23**	0.01	0.93	**0.07**
		320/80	0.68	**0.31**	0.01	1.00	–	0.94	**0.06**	–	0.94	**0.06**
		800/200	0.47	**0.50**	0.03	1.00	–	0.97	**0.03**	–	0.97	**0.03**
	Large	80/20	0.18	**0.76**	0.07	0.97	**0.03**	0.04	**0.78**	0.18	0.59	**0.41**[b]
		160/40	0.02	**0.92**	0.06	0.88	**0.12**	0.02	**0.92**	0.05	0.33	**0.67**[b]
		320/80	–	**0.94**	0.06	0.33	**0.67**	–	**0.99**	0.01	0.03	**0.97**
		800/200	–	**0.97**	0.03	–	**1.00**	–	**1.00**	–	–	**1.00**

The hypothesized correct enumeration rates are in bold. DIF, Differential item functioning or measurement noninvariance; AIC, Akaike information criterion; BIC, Bayesian information criterion; saBIC, sample-size adjusted BIC; HBIC, hierarchical BIC. Due to rounding 0.00 means one or two replications out of 500.

[a] *We compared one-, two-, and three-class models, but the number is not shown when the enumeration rates are zero across all conditions.*

[b] *The three-class model was selected with a small proportion.*

including factor mean differences, the AIC tended to over-extract latent classes more frequently.

The impact of mixing proportion (balanced and unbalanced) was not very noticeable and inconsistent across simulation conditions. For example, the unbalanced conditions showed slightly lower correct enumeration rates of BIC when there was only measurement noninvariance at the measurement model. It is possibly related to the bias in noninvariance size. That is, noninvariance size was underestimated more in the unbalanced conditions (see bias and relative bias below). However, although not very noticeable, the opposite pattern was observed in the conditions of both measurement noninvariance and factor mean difference conditions. Overall, accurate class enumeration (and relatedly accurate parameter estimation) appeared more challenging when one class had a notably small sample size under low class separation, but when sample size and class separation became larger with both measurement noninvariance and factor mean difference, the impact of a small class was not observed. However, it should be replicated in future research. When two classes were identified, the mixing proportions were generally well recovered with about 50%, 50% for balanced conditions and about 80%, 20% for unbalanced conditions save the unbalanced conditions under small measurement noninvariance in which the mixing

TABLE 4 | The class enumeration rates of second-order growth mixture modeling under the factor mean difference conditions.

DIF location	DIF size	Sample size	AIC		BIC		saBIC		HBIC	
			Number of latent classes							
			2	3	1	2[a]	2	3	1	2
Loading	Small	50/50	**0.80**	0.20	0.00	**1.00**	**0.44**	0.56	–	**0.98**[c]
		100/100	**0.82**	0.18	–	**1.00**	**0.88**	0.12	–	**1.00**[c]
		200/200	**0.84**	0.16	–	**1.00**	**0.98**	0.02	–	**1.00**
		500/500	**0.95**	0.05	–	**1.00**	**1.00**	–	–	**1.00**
	Large	50/50	**0.84**	0.16	–	**1.00**	**0.48**	0.52	–	**0.96**[c]
		100/100	**0.85**	0.15	–	**1.00**	**0.89**	0.11	–	**0.99**[c]
		200/200	**0.89**	0.11	–	**1.00**	**0.99**	0.01	–	**1.00**
		500/500	**0.91**	0.09	–	**1.00**	**1.00**	–	–	**1.00**
Intercept	Small	50/50	**0.77**[b]	0.21	0.59	**0.41**	**0.43**[b]	0.56	0.09	**0.89**[c]
		100/100	**0.78**[b]	0.22	0.02	**0.98**	**0.84**[b]	0.15	0.00	**1.00**
		200/200	**0.81**[b]	0.19	0.00	**1.00**	**0.98**[b]	0.02	0.00	**1.00**
		500/500	**0.94**	0.06	–	**1.00**	**1.00**	–	–	**1.00**
	Large	50/50	**0.77**	0.23	0.01	**0.99**	**0.40**	0.60	0.00	**0.98**[c]
		100/100	**0.81**	0.19	–	**1.00**	**0.86**	0.14	–	**1.00**[c]
		200/200	**0.85**	0.15	–	**1.00**	**0.99**	0.01	–	**1.00**
		500/500	**0.97**	0.03	–	**1.00**	**1.00**	–	–	**1.00**
Loading	Small	80/20	**0.80**[b]	0.19	0.00	**1.00**	**0.52**[b]	0.47	0.00	**0.99**[c]
		160/40	**0.85**	0.15	–	**1.00**	**0.88**	0.12	–	**1.00**[c]
		320/80	**0.83**	0.17	–	**1.00**	**0.99**	0.01	–	**1.00**
		800/200	**0.93**	0.07	–	**1.00**	**1.00**		–	**1.00**
	Large	80/20	**0.82**	0.18	–	**1.00**	**0.51**	0.49	–	**0.99**[c]
		160/40	**0.86**	0.14	–	**1.00**	**0.89**	0.11	–	**1.00**[c]
		320/80	**0.86**	0.14	–	**1.00**	**0.99**	0.01	–	**1.00**
		800/200	**0.92**	0.08	–	**1.00**	**1.00**	.	–	**1.00**
Intercept	Small	80/20	**0.77**[b]	0.17	0.06	**0.94**	**0.48**[b]	0.50	0.06	**0.94**[c]
		160/40	**0.80**[b]	0.17	0.05	**0.95**	**0.84**[b]	0.12	0.04	**.96**
		320/80	**0.84**[b]	0.15	0.01	**0.99**	**0.98**[b]	0.02	0.01	**.99**
		800/200	**0.88**	0.12	–	**1.00**	**1.00**	–	–	**1.00**
	Large	80/20	**0.80**	0.20	0.00	**1.00**	**0.45**	0.55	0.00	**0.99**[c]
		160/40	**0.84**	0.16	–	**1.00**	**0.88**	0.12	–	**1.00**[c]
		320/80	**0.87**	0.13	–	**1.00**	**0.99**	0.01	–	**1.00**
		800/200	**0.88**	0.12	–	**1.00**	**1.00**	–	–	**1.00**

The hypothesized correct enumeration rates are in bold. DIF, Differential item functioning or measurement noninvariance; AIC, Akaike information criterion; BIC, Bayesian information criterion; saBIC, sample-size adjusted BIC; HBIC, hierarchical BIC. Due to rounding 0.00 means one or two replications out of 500.

[a] *We compared one-, two-, and three-class models, but the three-class model was not selected across all conditions.*

[b] *The one-class model was selected with a small proportion.*

[c] *The three-class model was selected with a small proportion.*

proportions turned out to be close to 50%, 50% as sample size became smaller.

Bias and Relative Bias

We examined the bias or relative bias of parameter estimates in terms of the size of noninvariance and factor mean differences when two latent classes were correctly identified. Because SOGMM was correctly specified, all parameter estimates were expected to be unbiased. As expected, the relative bias of factor loading and intercept noninvariance was negligible in most conditions when there were differences in both measurement parameters and factor means between classes. On the contrary, when measurement noninvariance was the only sources of population heterogeneity between classes, the noninvariance size was underestimated consistently across conditions as shown in **Table 5** (left panel). The relative bias ranged from −0.788 to −0.010. The variability of relative bias was in general associated with sample size (the larger, the smaller), mixing proportion (smaller with the balanced proportions), and noninvariance size (smaller with large noninvariance). Note that under these conditions growth parameters were still unbiased. It appeared that the class specific measurement parameters

TABLE 5 | The bias and relative bias of the parameter estimates in second-order growth mixture modeling.

DIF location	DIF size	Sample size	No difference	Difference	No difference		Difference	
			Raw bias	Rel. bias	Raw bias		Rel. bias	
			DIF	DIF	Intercept	Slope	Intercept *d*	Slope *d*
Loading	Small	50/50	−0.600	0.022	−0.039	0.121	−0.008	−0.020
		100/100	−0.515	0.008	−0.021	0.053	0.007	−0.013
		200/200	−0.403	−0.005	−0.005	0.027	0.000	−0.005
		500/500	−0.190	0.000	−0.003	0.007	−0.001	−0.005
	Large	50/50	−0.425	0.000	−0.020	0.027	−0.009	−0.023
		100/100	−0.321	0.000	−0.009	0.012	0.001	−0.003
		200/200	−0.184	−0.003	−0.002	0.008	0.000	0.000
		500/500	−0.018	0.003	−0.001	0.000	0.001	0.000
Intercept	Small	50/50	−0.157	−0.077	−0.129	0.230	−0.034	−0.095
		100/100	−0.327	0.000	−0.111	0.141	−0.018	−0.018
		200/200	−0.267	0.005	−0.060	0.095	−0.001	−0.005
		500/500	−0.173	0.000	−0.006	0.028	0.002	0.009
	Large	50/50	−0.310	−0.004	−0.026	0.049	0.002	0.010
		100/100	−0.231	−0.006	−0.012	0.014	0.000	0.000
		200/200	−0.091	−0.003	−0.001	0.005	0.004	0.000
		500/500	−0.010	0.008	0.004	0.001	−0.002	0.001
Loading	Small	80/20	−0.788	0.007	−0.105	0.210	0.013	−0.080
		160/40	−0.713	0.032	−0.041	0.119	0.002	−0.015
		320/80	−0.493	0.015	0.014	0.059	0.002	−0.003
		800/200	−0.240	−0.003	−0.004	0.012	0.002	−0.003
	Large	80/20	−0.456	0.003	−0.019	0.028	0.003	−0.005
		160/40	−0.450	0.006	−0.017	0.021	0.002	0.006
		320/80	−0.408	0.003	0.005	0.016	0.001	0.003
		800/200	−0.375	0.000	−0.003	0.005	0.001	0.001
Intercept	Small	80/20	−0.357	−0.142	−0.180	0.251	0.016	−0.143
		160/40	−0.395	−0.030	−0.122	0.210	−0.004	−0.005
		320/80	−0.317	−0.020	−0.100	0.154	0.004	0.006
		800/200	−0.328	−0.005	−0.074	0.073	0.001	−0.008
	Large	80/20	−0.463	−0.018	−0.091	0.100	0.012	−0.065
		160/40	−0.339	−0.003	−0.033	0.029	0.003	−0.013
		320/80	−0.304	−0.014	−0.003	0.013	0.004	−0.003
		800/200	−0.144	−0.003	0.001	0.003	0.001	−0.003

DIF, Differential item functioning or measurement non-invariance; No difference, no factor mean difference; Difference, factor mean difference; Rel. bias, relative bias; Intercept, intercept factor mean; Slope, slope factor mean; Intercept d, intercept factor mean difference; Slope d, slope factor mean difference.

(i.e., noninvariance) in the first order model were in general less accurately estimated than the class specific growth parameters (i.e., structural parameters) in the second order model when class separation was low. The estimation of the first (i.e., noninvariance size) improved with a larger sample and bigger separation. With respect to mixing proportions, the estimation could be less accurate with a disproportionately smaller class size in the unbalanced conditions when the total sample size was small (e.g., 20 when total $N = 100$)[1]. The underestimation of noninvariance size was possibly related to the lower enumeration rates under the measurement noninvariance only conditions (no factor mean difference) because in these conditions noninvariance was the only source of heterogeneity that separated two classes.

The right panel of **Table 5** presents the bias or relative bias of factor mean difference estimates between classes. Of note is that the bias or relative bias was estimated for the factor mean differences (not for the factor means). When there were factor mean differences, the estimated differences were generally unbiased regardless of simulation factors. The relative bias of

[1] It should be noted that the accuracy of class membership assignment appeared not related to the underestimation of noninvariance size because (a) the growth parameters were unbiased under the same conditions, and (b) we did not observe an apparent association between the assignment accuracy and bias. For example, under the unbalanced conditions in which the underestimation was larger, the assignment accuracy was even slightly higher compared to the balanced conditions.

the factor mean differences between classes was <0.05 across conditions except the smallest sample size conditions ($N = 100$; see the two columns of the last panel in **Table 5**). When there was no factor mean difference and two classes were different only due to measurement noninvariance, the estimated factor means generally showed no difference between classes (i.e., no bias) with large sample, but when sample size was small, bigger size of raw bias was observed (See the two columns of the middle panel in **Table 5**).

DISCUSSION

When researchers run GMM, it is a common practice to use composite scores of repeated measures to model the baseline performance and growth over time. This could be problematic when the measure does not have desirable psychometric properties because GMM does not allow evaluating measurement models. In this study we addressed one of these issues—measurement noninvariance. When there was measurement noninvariance between unknown groups, we investigated the impact of the ignored noninvariance on the performance of GMM, particularly, the accuracy of class enumeration and the parameter recovery. In addition, we examined the performance of SOGMM that incorporates measurement models and allows measurement noninvariance between latent classes.

First, we hypothesized that due to unmodeled noninvariance in items GMM would incorrectly identify two latent classes showing differences in factor means between classes when there was no difference in factor means. In Study 1, this hypothesis was not supported because the BIC and HBIC mostly selected a one-class model as a best-fitting model. However, this finding should not be interpreted as no impact of the ignored measurement noninvariance on the GMM class enumeration. Rather, it might indicate that overall, GMM is not very sensitive to a small degree of population heterogeneity. This was confirmed in the conditions with both measurement noninvariance and factor mean differences. Although the generated class separation (MD or mahalanobis distance = 2) in the factor mean differences was considered large, the BIC, for example, supported one class more often when sample size was 400 or less. Under these conditions the enumeration rates of BIC were associated with the location and size of noninvariance, which implies that the ignored measurement noninvariance affected the performance of GMM, specifically, the enumeration accuracy of BIC and HBIC.

Second, as hypothesized, the parameter estimates of GMM, namely, the intercept and slope factor means were biased regardless of simulation factors. The location, size, and direction of bias were directly related to the location, size, and direction of unmodeled measurement noninvariance. That is, we observed positive bias in the intercept factor when positive noninvariance in the item intercepts were ignored and negative bias in the slope factor when negative noninvariance in the item factor loadings were ignored. When the size of noninvariance was doubled, the size of bias was also doubled. This finding is consistent to what Kim and Willson (2014a) found with multiple group LGM. Because GMM yields biased parameter estimates even if the number of latent classes is correctly detected, GMM is not recommended in the presence of measurement noninvariance.

Third, with respect to SOGMM, our hypothesis about high accuracy of class enumeration of SOGMM was partly supported in Study 2 because class enumeration rates largely depended on class separation and sample size. When the class separation was large under the conditions of both measurement noninvariance and factor mean differences, the correct enumeration rates of BIC were almost 100% even with a very small sample size (i.e., 100). However, when the class separation was low (small noninvariance only) and sample size was small (400 or less), the correct enumeration rates of BIC dropped substantially (e.g., 0%). A previous simulation study (Lubke and Neale, 2008) that investigated the class enumeration rates in detecting measurement noninvariance also found that more parsimonious models (e.g., one-class model) were favored indicating no measurement noninvariance. The overall insensitivity of ICs to the presence of small measurement noninvariance between latent classes can be explained as relatively low class separation that measurement noninvariance created. The small noninvariance in the intercepts corresponded to MD = 1, which is considered as small class separation in the literature. It is widely recognized that class separation is greatly related to the accuracy of class enumeration (e.g., Henson et al., 2007; Tofighi and Enders, 2008; Chen et al., 2010).

Fourth, the hypothesis that SOGMM would yield unbiased estimates was also partly correct. As hypothesized, the intercept and slope factor means of SOGMM were generally unbiased. When sample size was very small, we observed some biased estimates of these parameters. On the other hand, the size of noninvariance in the intercepts and factor loadings were generally underestimated. This underestimation of noninvariance size might make it more difficult for ICs to detect the difference and be partly related to the low class enumeration accuracy when measurement noninvariance was the only source of population heterogeneity between classes.

With respect to information criteria, the findings in this study generally conform to those of previous studies. The BIC showed excellent performance in identifying the number of classes when class separation was large and sample size was large (Nylund et al., 2007; Lubke and Neale, 2008; Li et al., 2009). When both class separation was low and sample size was small, the BIC tended to under-extract latent classes (Kim et al., 2016). The HBIC showed similar or slightly better performance than the BIC. The outperformance of HBIC was prominent when sample size and class separation were small, which is consistent to the findings of previous studies (e.g., Zhao et al., 2015). It should be noted that the over-extraction of latent classes was also observed with the HBIC, which is possibly due to under-penalization of model complexity compared to the BIC although the over-extraction was not very serious in this study. The saBIC showed more consistent performance across simulation conditions, but its accuracy was lower compared to BIC and HBIC when these two ICs worked reasonably. The AIC seemed least affected by simulation factors usually showing consistent enumeration rates across simulation conditions and most sensitive to population heterogeneity in the extreme conditions (i.e., smallest sample size in this study under low class separation; e.g., Lukočienė and Vermunt, 2010; Lukočienė et al., 2010; Kim et al., 2016), but the performance of AIC was generally not optimal and also tended to over-extract latent classes (e.g., Bozdogan, 1987; Nylund et al., 2007; Tein et al., 2013). As explained in previous studies (Henson et al., 2007; Kim et al., 2015, 2016), the BIC uses the natural logarithm of sample size multiplied by the number of free parameters (k) to penalize for model complexity. The penalty of BIC on additionally estimated parameters (i.e., additional latent class) is more severe than that of AIC (2^*k). Thus, when class separation is low, a complex model with more parameters from additional latent classes may not be favored with the BIC due to too severe penalty on model complexity relative to small differences between latent classes. Under these circumstances, the AIC as well as HBIC generally outperformed the BIC. Also the AIC is not supposed to be affected by sample size as much as the other ICs that include sample size in their computations.

Taken all together, when sample size is large (over 400 or 1,000 in this study) or class separation is expected to be large, the BIC or HBIC is recommended in GMM. When sample size is 400 or less and class separation is expected to be low, the saBIC seems a better choice in GMM. In SOGMM, if class separation is substantially large (MD = 2 or larger), the BIC or HBIC can be considered for class enumeration regardless of sample size. However, similar to GMM, when class separation is expected to be low and sample size is 400 or less, the saBIC is more recommended than the BIC and HBIC in determining the number of latent classes. The AIC could be a choice only when sample size is extremely small (100), but the mixture modeling is not recommended with this small sample.

Based on the findings in this study, it can be said that overall, GMM and SOGMM require large sample to correctly identify the number of classes and yield unbiased parameter estimates (Tueller and Lubke, 2010; Depaoli, 2013; Li and Harring, 2016). Vermunt (2010) noted that sample size 500 can be considered small for correct class enumeration especially under poor class separation, which was observed in this study particularly with the BIC. Even when the model is correctly specified as demonstrated in Study 2 with SOGMM, latent classes are not expected to be properly detected with small samples. Even when the number of latent classes is correctly identified, the parameter estimates could be substantially biased. Therefore, researchers interested in GMM or SOGMM should consider a large sample. This study also confirmed that class separation and sample size are generally major factors related to the class enumeration accuracy, which was consistently shown in the mixture modeling literature (e.g., Dias, 2004; Henson et al., 2007; Lubke and Neale, 2008; Tofighi and Enders, 2008; Chen et al., 2010).

We recommend SOGMM over GMM whenever possible for two major reasons. First, across all simulation conditions SOGMM produced unbiased estimates of growth trajectory parameters which are generally the focal interest of growth analysis in psychological research. SOGMM is advantageous because it includes measurement models of repeated measures that take into account measurement error and allows heterogeneity in measurement parameters between unknown groups. If there are differences in measurement models between potential groups, the differences can be captured by heterogeneous latent classes in SOGMM. As illustrated in this study, when the heterogeneity in measurement models is ignored and GMM is run, the parameter estimates of GMM are expected to be biased to the degree of the size of ignored differences. For example, developmental psychologists may observe an inflated difference between increasing and decreasing trajectory classes in terms of depressive symptoms. Or they may observe a smaller difference between them due to biased trajectory estimates. It was also observed that the ignored measurement noninvariance impacted on the enumeration rates of BIC and HBIC. Of note is that measurement invariance across latent classes cannot be tested separately using longitudinal common factor models because latent classes are unknown in advance. Thus, SOGMM is more imperative when researchers are interested in unknown clustering of growth trajectories. Second, SOGMM generally showed more accurate class enumeration possibly because it could directly detect any differences in the measurement models as well as in the growth model.

Finally, it should be kept in mind that some simulation factors were manipulated for the purpose of the study and hence generalization of the results beyond the simulation settings should be done with caution. For example, to highlight the impact of ignored measurement noninvariance between latent classes, we assumed measurement invariance over time. In reality, this assumption is not guaranteed and researchers should test and establish the temporal measurement invariance. Another assumption we made for the simplicity of discussion is no error correlation over time, but this assumption is less likely to be met with real data. In the presence of error correlation over time, the correct class enumeration is possibly more challenging because class separation becomes lower. Future research is called for the impact of different types of error structures on the performance of GMM and SOGMM.

AUTHOR CONTRIBUTIONS

All authors (EK, YW) contribute to the paper substantially and agree to be accountable for the content of the work.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpsyg.2017.01499/full#supplementary-material

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Contr.* 19, 716–723. doi: 10.1109/TAC.1974.1100705

Baams, L., Overbeek, G., Dubas, J. S., and van Aken, M. A. G. (2014). On early starters and late bloomers: the development of sexual behavior in adolescence across personality types. *J. Sex Res.* 51, 754–764. doi: 10.1080/00224499.2013.802758

Bauer, D. J. (2007). Observations on the use of growth mixture models in psychological research. *Multivariate Behav. Res.* 42, 757–786. doi: 10.1080/00273170701710338

Bauer, D. J., and Curran, P. J. (2003). Distributional assumptions of growth mixture models: implications for overextraction of latent trajectory classes. *Psychol. Methods* 8, 338–363. doi: 10.1037/1082-989X.8.3.338

Bozdogan, H. (1987). Model selection and akaike's information criterion (AIC): the general theory and its analytical extensions. *Psychometrika* 52, 345–370. doi: 10.1007/BF02294361

Brendgen, M., Girard, A., Vitaro, F., Dionne, G., and Boivin, M. (2016). Personal and familial predictors of peer victimization trajectories from primary to secondary school. *Dev. Psychol.* 52, 1103–1114. doi: 10.1037/dev0000107

Cabrera, O. A., Adler, A. B., and Bliese, P. D. (2016). Growth mixture modeling of post-combat aggression: application to soldiers deployed to Iraq. *Psychiatry Res.* 246, 539–544. doi: 10.1016/j.psychres.2016.10.035

Celeux, G., and Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *J. Classif.* 13, 195–212. doi: 10.1007/BF01246098

Chen, Q., Kwok, O.-M., Luo, W., and Willson, V. L. (2010). The impact of ignoring a level of nesting structure in multilevel growth mixture models: a Monte Carlo study. *Struct. Equ. Model. Multidiscipl. J.* 17, 570–589. doi: 10.1080/10705511.2010.510046

Depaoli, S. (2013). Mixture class recovery in GMM under varying degrees of class separation: Frequentist versus Bayesian estimation. *Psychol. Methods* 18, 186–219. doi: 10.1037/a0031609

Dias, J. G. (2004). *Finite Mixture Models*. Review, *Applications, and Computer-Intensive Methods*. Ridderkerk: Ridderprint.

Feldman, B. J., Masyn, K. E., and Conger, R. D. (2009). New approaches to studying problem behaviors: a comparison of methods for modeling longitudinal, categorical adolescent drinking data. *Dev. Psychol.* 45, 652–676. doi: 10.1037/a0014851

Frankfurt, S., Frazier, P., Syed, M., and Jung, K. R. (2016). Using group-based trajectory and growth mixture modeling to identify classes of change trajectories. *Couns. Psychol.* 44, 622–660. doi: 10.1177/0011000016658097

Gollini, I., and Murphy, T. B. (2014). Mixture of latent trait analyzers for model-based clustering of categorical data. *Stat. Comput.* 24, 569–588. doi: 10.1007/s11222-013-9389-1

Grimm, K. J., and Ram, N. (2009). A second-order growth mixture model for developmental research. *Res. Hum. Dev.* 6, 121–143. doi: 10.1080/15427600902911221

Hayward, R. D., and Krause, N. (2016). Classes of individual growth trajectories of religious coping in older adulthood: patterns and predictors. *Res. Aging* 38, 554–579. doi: 10.1177/0164027515593347

Henson, J. M., Reise, S. P., and Kim, K. H. (2007). Detecting mixtures from structural model differences using latent variable mixture modeling: a comparison of relative model fit statistics. *Struct. Equ. Model. Multidiscipl. J.* 14, 202–226. doi: 10.1080/10705510709336744

Hill, R. M., Mellick, W., Temple, J. R., and Sharp, C. (2017). The role of bullying in depressive symptoms from adolescence to emerging adulthood: a growth mixture model. *J. Affect. Disord.* 207, 1–8. doi: 10.1016/j.jad.2016.09.007

Hoogland, J. J., and Boomsma, A. (1998). Robustness studies in covariance structure modeling: an overview and a meta-analysis. *Sociol. Methods Res.* 26, 329–367. doi: 10.1177/0049124198026003003

Jak, S., Oort, F. J., and Dolan, C. V. (2014). Measurement bias in multilevel data. *Struct. Equ. Model. Multidiscipl. J.* 21, 31–39. doi: 10.1080/10705511.2014.856694

Kim, E. S., and Willson, V. L. (2014a). Measurement invariance across groups in latent growth modeling. *Struct. Equ. Model. Multidiscipl. J.* 21, 408–424. doi: 10.1080/10705511.2014.915374

Kim, E. S., and Willson, V. L. (2014b). Testing measurement invariance across groups in longitudinal data: multigroup second-order latent growth model. *Struct. Equ. Model. Multidiscipl. J.* 21, 566–576. doi: 10.1080/10705511.2014.919821

Kim, E. S., Joo, S.-H., Lee, P., Wang, Y., and Stark, S. (2016). Measurement invariance testing across between-level latent classes using multilevel factor mixture modeling. *Struct. Equ. Model. Multidiscipl. J.* 23, 870–887. doi: 10.1080/10705511.2016.1196108

Kim, E. S., Yoon, M., and Lee, T. (2012). Testing measurement invariance using MIMIC: likelihood ratio test with a critical value adjustment. *Educ. Psychol. Meas.* 72, 469–492. doi: 10.1177/0013164411427395

Kim, E. S., Yoon, M., Wen, Y., Luo, W., and Kwok, O. (2015). Within-level group factorial invariance with multilevel data: multilevel factor mixture and multilevel MIMIC models. *Struct. Equ. Model. Multidiscipl. J.* 22, 603–616. doi: 10.1080/10705511.2014.938217

Lee, T. K., Wickrama, K. A. S., O'Neal, C. W., and Lorenz, F. O. (2017). Social stratification of general psychopathology trajectories and young adult social outcomes: a second-order growth mixture analysis over the early life course. *J. Affect. Disord.* 208, 375–383. doi: 10.1016/j.jad.2016.08.037

Leite, W. L. (2007). A comparison of latent growth models for constructs measured by multiple items. *Struct. Equ. Model. Multidiscipl. J.* 14, 581–610. doi: 10.1080/10705510701575438

Li, F., Cohen, A. S., Kim, S.-H., and Cho, S.-J. (2009). Model selection methods for mixture dichotomous IRT models. *Appl. Psychol. Meas.* 33, 353–373. doi: 10.1177/0146621608326422

Li, F., Duncan, T. E., and Hops, H. (2001). Examining developmental trajectories in adolescent alcohol use using piecewise growth mixture modeling analysis. *J. Stud. Alcohol* 62, 199–210. doi: 10.15288/jsa.2001.62.199

Li, M., and Harring, J. R. (2016). Investigating approaches to estimating covariate effects in growth mixture modeling: a simulation study. *Educ. Psychol. Meas.* doi: 10.1177/0013164416653789. [Epub ahead of print].

Lubke, G. H., and Muthén, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychol. Methods* 10, 21–39. doi: 10.1037/1082-989X.10.1.21

Lubke, G., and Neale, M. (2008). Distinguishing between latent classes and continuous factors with categorical outcomes: class invariance of parameters of factor mixture models. *Multivariate Behav. Res.* 43, 592–620. doi: 10.1080/00273170802490673

Lubke, G., and Neale, M. C. (2006). Distinguishing between latent classes and continuous factors: resolution by maximum likelihood? *Multivariate Behav. Res.* 41, 499–532. doi: 10.1207/s15327906mbr4104_4

Lukočienė, O., and Vermunt, J. K. (2010). "Determining the number of components in mixture models for hierarchical data," in *Advances in Data Analysis, Data Handling and Business Intelligence,* eds A. Fink, L. Berthold, W. Seidel, and A. Ultsch (Berlin: Springer), 241–249.

Lukočienė, O., Varriale, R., and Vermunt, J. K. (2010). The simultaneous decision(s) about the number of lower- and higher-level classes in multilevel latent class analysis. *Sociol. Methodol.* 40, 247–283. doi: 10.1111/j.1467-9531.2010.01231.x

McArdle, J. J. (1988). "Dynamic but structural equation modeling of repeated measures data," in *Handbook of Multivariate Experimental Psychology*, eds J. R. Nesselroade and R. B. Cattell (New York, NY: Plenum), 561–614.

McLachlan, G., and Peel, D. (2000). *Finite Mixture Models*. Hoboken, NJ: Wiley.

Meredith, W., and Tisak, J. (1990). Latent curve analysis. *Psychometrika* 55, 107–122. doi: 10.1007/BF02294746

Millsap, R. E., and Kwok, O.-M. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychol. Methods* 9, 93–115. doi: 10.1037/1082-989X.9.1.93

Miner, J. L., and Clarke-Stewart, K. A. (2008). Trajectories of externalizing behavior from age 2 to age 9: relations with gender, temperament, ethnicity, parenting, and rater. *Dev. Psychol.* 44, 771–786. doi: 10.1037/0012-1649.44.3.771

Muthén, B. (2004). "Latent variable analysis: growth mixture modeling and related techniques for longitudinal data," in *The Sage Handbook of Quantitative Methodology for the Social Sciences,* ed D. Kaplan (Newbury Park, CA: Sage), 345–368.

Muthén, B. O., and Muthén, L. K. (2014). *Mplus 7.3 [Computer software]*. Los Angenles, CA: Muthén & Muthén.

Muthén, B., Khoo, S. T., Francis, D. J., and Boscardin, C. K. (2000). "Analysis of reading skills development from kindergarten through first grade: an application of growth mixture modeling to sequential processes," in *Multilevel Modeling: Methodological Advances, Issues and Applications,* eds S. R. Reise and N. Duan (Mahwah, NJ: Lawrence Erlbaum Associates, Inc.), 71–89.

Muthén, L. K., and Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Struct. Equ. Model. Multidiscipl. J.* 9, 599–620. doi: 10.1207/S15328007SEM0904_8

Nash, W. P., Boasso, A. M., Steenkamp, M. M., Larson, J. L., Lubin, R. E., and Litz, B. T. (2015). Posttraumatic stress in deployed marines: prospective trajectories of early adaptation. *J. Abnorm. Psychol.* 124, 155–171. doi: 10.1037/abn0000020

Nylund, K. L., Asparouhov, T., and Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: a Monte Carlo simulation study. *Struct. Equ. Model. Multidiscipl. J.* 14, 535–569. doi: 10.1080/10705510701575396

Oshri, A., Carlson, M. W., Kwon, J. A., Zeichner, A., and Wickrama, K. K. A. S. (2017). Developmental growth trajectories of self-esteem in adolescence: associations with child neglect and drug use and abuse in young adulthood. *J. Youth Adolesc.* 46, 151–164. doi: 10.1007/s10964-016-0483-5

Raykov, T., Marcoulides, G. A., and Li, C.-H. (2012). Measurement invariance for latent constructs in multiple populations: a critical view and refocus. *Educ. Psychol. Meas.* 72, 954–974. doi: 10.1177/0013164412441607

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.* 6, 461–464. doi: 10.1214/aos/1176344136

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika* 52, 333–343. doi: 10.1007/BF02294360

Stark, S., Chernyshenko, O. S., and Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: towad a unified strategy. *J. Appl. Psychol.* 91, 1292–1306. doi: 10.1037/0021-9010.91.6.1292

Tay, L., Newman, D. A., and Vermunt, J. K. (2011). Using mixed-measurement item response theory with covariates (MM-IRT-C) to ascertain observed and unobserved measurement equivalence. *Organ. Res. Methods* 14, 147–176. doi: 10.1177/1094428110366037

Tein, J., Coxe, S., and Cham, H. (2013). Statistical power to detect the correct number of classes in latent profile analysis. *Struct. Equ. Model. Multidiscipl. J.* 20, 640–657. doi: 10.1080/10705511.2013.824781

Tofighi, D., and Enders, C. K. (2008). "Identifying the correct number of classes in growth mixture models," in *Advances in Latent Variable Mixture Models,* eds G. R. Hancock and K. M. Samuelsen (Charlotte, NC: Information Age Publishing, Inc.), 317–341.

Tueller, S., and Lubke, G. (2010). Evaluation of structural equation mixture models: parameter estimates and correct class assignment. *Struct. Equ. Model. Multidiscipl. J.* 17, 165–192. doi: 10.1080/10705511003659318

Vermunt, J. K. (2010). Latent class modeling with covariates: two improved three-step approaches. *Polit. Anal.* 18, 450–469. doi: 10.1093/pan/mpq025

Vermunt, J. K. (2011). K-means may perform as well as mixture model clustering but may also be much worse: comment on Steinley and Brusco (2011). *Psychol. Methods* 16, 82–88. doi: 10.1037/a0020144

Wang, Y. L., Chan, H.-Y., Lin, C.-W., and Li, J.-R. (2015). Association of parental warmth and harsh discipline with developmental trajectories of depressive symptoms among adolescents in Chinese society. *J. Fam. Psychol.* 29, 895–906. doi: 10.1037/a0039505

Wirth, R. J. (2009). *The Effects of Measurement Non-invariance on Parameter Estimation in Latent Growth Models*. University of North Carolina, Chapel Hill, NC. Retrieved from ProQuest Dissertations and Theses (AAI 3331053).

Wu, W., West, S. G., and Taylor, A. B. (2009). Evaluating model fit for growth curve models: integration of fit indices from SEM and MLM frameworks. *Psychol. Methods* 14, 183–201. doi: 10.1037/a0015858

Zhao, J. (2014). Efficient model selection for mixtures of probabilistic PCA via hierarchical BIC. *IEEE Trans. Cybern.* 44, 1871–1883. doi: 10.1109/TCYB.2014.2298401

Zhao, J., Jin, L., and Shi, L. (2015). Mixture model selection via hierarchical BIC. *Comput. Stat. Data Anal.* 88, 139–153. doi: 10.1016/j.csda.2015.01.019

Zhao, J., Yu, P. L., and Shi, L. (2013). *Model Selection for Mixtures of Factor Analyzers via Hierarchical BIC*. Yunnan: School of Statistics and Mathematics, Yunnan University of Finance and Economics.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer DM and handling Editor declared their shared affiliation.

frontiers in Psychology

ORIGINAL RESEARCH
published: 05 June 2018
doi: 10.3389/fpsyg.2018.00790

Analyzing Complex Longitudinal Data in Educational Research: A Demonstration With Project English Language and Literacy Acquisition (ELLA) Data Using xxM

Oi-Man Kwok[1,2]*, *Mark Hok-Chio Lai*[3], *Fuhui Tong*[1,2], *Rafael Lara-Alecio*[1,2], *Beverly Irby*[1,2,4], *Myeongsun Yoon*[1,2] and *Yu-Chen Yeh*[1]

[1] Department of Educational Psychology, Texas A&M University, College Station, TX, United States, [2] Center for Research & Development in Dual Language & Literacy Acquisition (CRDLLA), College Station, TX, United States, [3] School of Education, University of Cincinnati, Cincinnati, OH, United States, [4] Department of Educational Administration and Human Resource Development, Education Leadership Research Center, Texas A&M University, College Station, TX, United States

OPEN ACCESS

Edited by:
Holmes Finch,
Ball State University, United States

Reviewed by:
Akihito Kamata,
Southern Methodist University,
United States
Ke-Hai Yuan,
University of Notre Dame,
United States

***Correspondence:**
Oi-Man Kwok
omkwok@email.tamu.edu

Specialty section:
This article was submitted to
Quantitative Psychology and
Measurement,
a section of the journal
Frontiers in Psychology

Received: 31 July 2017
Accepted: 03 May 2018
Published: 05 June 2018

Citation:
Kwok O-M, Lai MH-C, Tong F,
Lara-Alecio R, Irby B, Yoon M and Yeh
Y-C (2018) Analyzing Complex
Longitudinal Data in Educational
Research: A Demonstration With
Project English Language and Literacy
Acquisition (ELLA) Data Using xxM.
Front. Psychol. 9:790.
doi: 10.3389/fpsyg.2018.00790

When analyzing complex longitudinal data, especially data from different educational settings, researchers generally focus only on the mean part (i.e., the regression coefficients), ignoring the equally important random part (i.e., the random effect variances) of the model. By using Project English Language and Literacy Acquisition (ELLA) data, we demonstrated the importance of taking the complex data structure into account by carefully specifying the random part of the model, showing that not only can it affect the variance estimates, the standard errors, and the tests of significance of the regression coefficients, it also can offer different perspectives of the data, such as information related to the developmental process. We used xxM (Mehta, 2013), which can flexibly estimate different grade-level variances separately and the potential carryover effect from each grade factor to the later time measures. Implications of the findings and limitations of the study are discussed.

Keywords: longitudinal data analysis, multilevel structural equation models, educational psychology, intervention, bilingual education

INTRODUCTION

Educational researchers have always involved complex data structure. For example, in cross-sectional studies, students are likely nested within classrooms and schools at a particular time point (i.e., a strictly hierarchical structure), and while they may come from different neighborhoods, neighborhoods and schools are not nested but crossed with each other (i.e., a cross-classified structure). Similarly, for longitudinal data, repeated measures (e.g., reading achievement test scores collected at different grade levels from the same student) are nested within students while the students are likely to change classrooms over the course of study. A change of classroom results in a non-strictly hierarchical, but cross-classified structure, with repeated measures now nested within both students and classrooms, while students and classrooms are crossed with each other (see **Figure 1A**). Without adequately taking into account all these complex data structures, educational researchers not only may obtain biased parameter estimates and standard errors, but also they miss the opportunity to uncover important phenomena from their data.

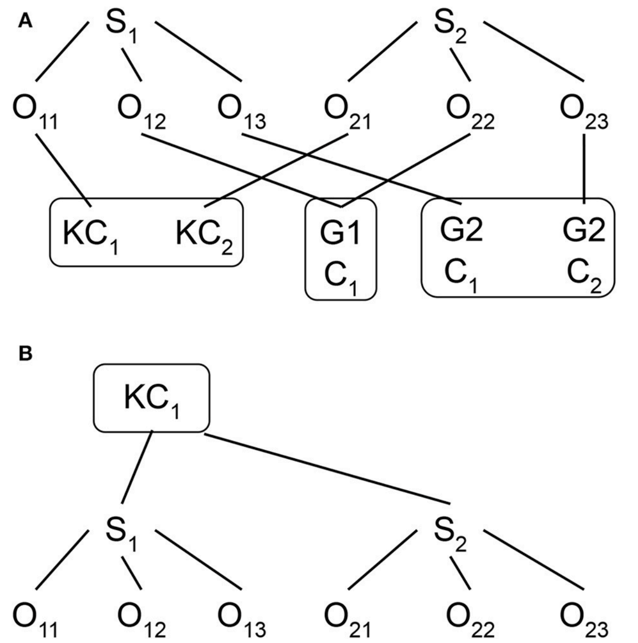

A
S_1
S_2
O_{11} O_{12} O_{13} O_{21} O_{22} O_{23}
KC_1 KC_2
G1 C_1
G2 C_1
G2 C_2
B
KC_1
S_1
S_2
O_{11} O_{12} O_{13} O_{21} O_{22} O_{23}

FIGURE 1 | (A) Model 2 data structure with repeated measures cross-classified by students and classrooms. O, Observation; S, Student. KC, Kindergarten classroom; G1, Grade 1; C1, Classroom 1; G2, Grade 2; C2, Classroom 2. **(B)** Model 1 data structure with repeated measures nested within students in kindergarten classrooms.

Although most educational researchers realize the importance of taking into account the complex data structure when they analyze their data, they may not be aware of how to *fully* address the complex data structure in their analysis and, as a result, they may only *partially* take into account the data structure. For instance, researchers may analyze the cross-classified data structure (e.g., repeated measures nested within students and classrooms, as presented in **Figure 1A**) by treating it as a strictly hierarchical data structure with the exclusion of the non-kindergarten classroom effect (e.g., first and second grade), as presented in **Figure 1B**. Without fully addressing the complex data structure, this mis-specified model may lead to a biased estimation of both fixed and random parameters and to incorrect significance tests for the parameter estimates (Meyers and Beretvas, 2006; Luo and Kwok, 2009).

The purpose of this paper was to demonstrate how to analyze this type of complex data structure with the use of data from the Project English Language and Literacy Acquisition (ELLA), a large-scale longitudinal study. The researchers intervened with and followed English language learners (ELLs) from kindergarten to third grade, which was funded by the U.S. Department of Education (Grant Number: R305P030032).

We first provide a brief review of the Project ELLA and the data derived from it. We, then, analyze the data with the commonly used hierarchical linear model [HLM] approach. We subsequently move from this HLM model to the more complex cross-classified random effect model (CCREM) which addresses the complex data structure issue by taking into account the classroom effect. However, the CCREM has its own limitations and is unable to address some of the important features of longitudinal data (which is representative of the dataset from Project ELLA), such as the potential carryover effect (i.e., the effect from the previous grade level on the later time measures). To address this special feature, we used the xxM software (Mehta, 2013; may be downloaded from http://xxm.times.uh.edu/), which could flexibly model the carryover effect during the analysis (the corresponding annotated input syntax and outputs are presented in the appendices). Finally, we discuss the implications of the different results based on different models and re-emphasize the importance of taking the carryover effect into account, followed the limitations of the study and directions for future research.

Project English Language and Literacy Acquisition (ELLA)

Project ELLA (Lara-Alecio, 2003) was a longitudinal, field-based, large-scale, experimental research project following the same group of native Spanish-speaking, English language learners (ELLs) over time (from kindergarten to third grade) in an urban school district in Southeast Texas. For more than 45% of the students in the district, Spanish was their first language was Spanish. The majority of students qualified for free or reduced-price lunch. All the materials and protocols of Project ELLA were approved by Institutional Review Board (IRB) at Texas A&M University.

Texas state law (Texas Education Code, 1995) has prohibited random selection and assignment to specific instructional delivery models in schools on the basis of individual students; therefore, the research team selected schools where structured English immersion (SEI) and/or transitional bilingual education (TBE) were being implemented within the target school district, and they randomly assigned the selected schools to either a control (typical practice) or an experimental (enhanced practice) setting. Hence, in the overall project, the researchers used an experimental design at the school (classroom) level and a quasi-experimental design with target learning outcomes at the student level.

In the current study, we used a partial data set from the original data. This data set included scores on the English version of the Woodcock Language Proficiency Battery–Picture Vocabulary subtest (EWPV) of 876 students at five time points: Time 1 = beginning of kindergarten (2004), Time 2 = end of kindergarten (2005), Time 3 = end of first grade (2006), Time 4 = end of second grade (2007), and Time 5 = end of third grade (2008).

As shown in **Table 1**, at Time 1, the study contained 24 schools with 56 classrooms and 876 students (46.00% of females and 53.65% of males) between the ages of 49 and 80 months ($M =$ 59.72 and $SD = 5.08$); EWPV data were available for 791 students. At Time 2, it contained 24 schools with 56 classrooms, with EWPV data available for 875 students (45.94% of females and 53.71% of males) between the ages of 61 and 92 months ($M = 71.72$ and $SD = 5.08$). At Time 3, it contained 24 schools with 54 classrooms, with EWPV data available for 643 students (46.19% of females and 53.34% of males) between the ages of

TABLE 1 | Descriptive statistics.

	Time 1 (*N* = 876)	Time 2 (*N* = 875)	Time 3 (*N* = 643)	Time 4 (*N* = 440)	Time 5 (*N* = 373)
Variables	***N* (%)/ *M(SD)***	***N* (%)/ *M(SD)***	***N* (%)/ *M(SD)***	***N* (%)/ *M(SD)***	***N* (%)/ *M(SD)***
Gender					
Male	470 (53.65%)	470 (53.71%)	343 (53.34%)	231 (52.50%)	191 (51.21%)
Female	403 (46.00%)	402 (45.94%)	297 (46.19%)	206 (46.82%)	179 (47.99%)
Age (months)	59.72 (5.08)	71.72 (5.08)	83.84 (5.01)	95.67 (4.61)	107.92 (4.64)
Conditions					
Control	390 (44.52%)	390 (44.57%)	295 (45.88%)	222 (50.45%)	192 (51.47%)
Treatment	486 (55.48%)	485 (55.43%)	348 (54.12%)	218 (49.55%)	181 (48.53%)

Time 1, beginning of kindergarten; Time 2, end of kindergarten; Time 3, end of first grade; Time 4, end of second grade; Time 5, end of third grade.

73 and 104 months ($M = 83.84$ and $SD = 5.01$). At Time 4, it contained 21 schools with 53 classrooms, with EWPV data available for 440 students (46.82% of females and 52.50% of males) between the ages of 85 and 112 months ($M = 95.67$ and $SD = 4.61$) had data on EWPV. At Time 5, it contained 21 schools with 60 classrooms, with EWPV data available for 373 students (47.99% of females and 51.21% of males) between the ages of 97 and 124 months ($M = 107.92$ and $SD = 4.64$).

Ways to Analyze Complex Longitudinal Data in Educational Research

We present three models, of which the first two are commonly used in educational research; namely, the hierarchical linear model (HLM) and the cross-classified random effect model (CCREM). The third, the xxM-UN1 model, is a more advanced and flexible model, which not only takes into account the complex data structure but also provides new modeling feature that allows researchers to examine such effects as potential carryover in longitudinal analysis. The results from these analytic approaches are compared, and the advantages and disadvantages of each model are discussed.

Even though the analyses have been conducted under both multilevel modeling (MLM; i.e., hierarchical linear modeling, HLM) and structural equation modeling (SEM) frameworks, we prefer using the multilevel modeling framework to present the models for our analyses, given its simplicity for comprehension and the equivalence between the two models (Curran, 2003; Bollen and Curran, 2006). For example, the average trend information in MLM is captured by the corresponding time-related latent factors (i.e., the means and variances of these latent factors) whereas the time-related information (i.e., the time frame of the study) is captured by the factor loadings between the time-related latent factors and the observed variables measured over time under the SEM framework. There are additional benefits of using SEM to analyze longitudinal data, including the availability of model fit indices and modification indices (Preacher et al., 2008; Kwok et al., 2010). Moreover, xxM (Mehta, 2013) provides a flexible framework for modeling complex multilevel and longitudinal data such as the carryover effect detailed later.

MODEL 1: THE TRADITIONAL THREE-LEVEL MULTILEVEL MODEL

Unlike the cross-sectional multilevel model, there is always an important predictor for longitudinal analysis: time. Researchers are particularly interested in examining the average trend of an outcome variable (in this paper, the Woodcock Language Proficiency Battery–Picture Vocabulary subtest; EWPV) over time. Nevertheless, many longitudinal and developmental phenomena are not linear in nature. In other words, the change of the outcome variable will not happen at a constant rate over time. For example, we may have a simple linear time-predicted model, Math = B0 + B1 Time + e, where Math is the math achievement outcome variable, Time is the time predictor with grade year as the unit, and e as the error. B0 is intercept, B1 (positive and significantly larger than zero) is the regression coefficient, which can be explained as one unit changes in time or one grade year passes, and B1 points change in the math achievement score. More importantly, this model implies the constant improvement in math achievement (with B1 points per grade year regardless of the actual grade year in which the students are located). Hence, fitting a nonlinear model rather than assuming a linear trend is common in analyzing longitudinal data (Kwok et al., 2010).

A relatively, more simple way to capture a nonlinear trend is using a piecewise model (Bryk and Raudenbush, 1992; Sayer and Willett, 1998; Snijders and Bosker, 1999; Duncan et al., 2006; Kwok et al., 2010). By dividing the nonlinear growth trend into different linear segments, one can easily understand the nonlinear trend by applying the same straightforward interpretation based on the simple linear growth rate coefficients. The key part of using the piecewise model is to determine how (many pieces) and where to divide the whole time frame into segments.

For our current demonstration, given the data collection time frame, we determined to use a piecewise model containing two pieces to capture the potential nonlinear trend, with the first piece containing the first two time measures (i.e., beginning and end of kindergarten) and the second piece containing the rest of the three time measures (i.e., end of first grade, end of second grade, and end of third grade). As described previously, we proposed analyzing the data with a piecewise model containing two pieces (a.k.a. a two-piece model). By using the traditional HLM, which assumes a strictly hierarchical structure, we have analyzed our data as a three-level model with repeated measures (level 1) nested within students (level 2) and students further nested within their corresponding kindergarten classrooms (level 3) without considering their mobility (i.e., change of classroom in later time points). The corresponding model equations are presented as follows:

Level 1 (repeated-measure level)

$$\mathbf{EWPV}_{tij} = \boldsymbol{\pi}_{0ij} + \boldsymbol{\pi}_{1ij}\ \mathbf{piece1}_{tij} + \boldsymbol{\pi}_{2ij}\ \mathbf{piece2}_{tij} + \boldsymbol{e}_{tij}, \quad (1)$$

where EWPV is the target outcome variable for the t-th repeated measure from the i-th student of the j-th ***kindergarten*** classroom, piece1 is the first time piece variable, which captures possible changes in EWPV in kindergarten, and piece2 is the second piece variable, which captures possible changes in EWPV from first to third grade.

We used the following coding scheme:

$$\left[\begin{array}{ccc} & piece1 & piece2 \\ K-begin & 0 & 0 \\ K-end & 1 & 0 \\ 1stGrade & 1 & 1 \\ 2ndGrade & 1 & 2 \\ 3rdGrade & 1 & 3 \end{array}\right],$$

with piece1 coded as (0,1,1,1,1) and piece2 coded as (0,0,1,2,3) for the five repeated measures. $\boldsymbol{\pi}_{0ij}$ is the intercept (or the baseline/predicted EWPV score at the beginning of kindergarten) based on the repeated measures from the i-th student of the j-th kindergarten classroom. Similarly, $\boldsymbol{\pi}_{1ij}$ is the linear rate of change of the first piece (i.e., from the beginning of kindergarten to the end of kindergarten) while $\boldsymbol{\pi}_{2ij}$ is the linear rate of change of the second piece (i.e., from the end of first to the end of third grade) from the i-th student of the j-th kindergarten classroom. Given that we had 876 students in the data, and we used the repeated measures from each student to fit the above two-piece model, we should have 876 sets of regression coefficients (i.e., $\boldsymbol{\pi}_{0ij}$, $\boldsymbol{\pi}_{1ij}$, & $\boldsymbol{\pi}_{2ij}$), which can be written into the following equations:

Level 2 (student level)

$$\boldsymbol{\pi}_{\mathbf{0ij}} = \boldsymbol{\beta}_{\mathbf{0j}} + \boldsymbol{u}_{\boldsymbol{0ij}} \quad (2)$$

$$\boldsymbol{\pi}_{\mathbf{1ij}} = \boldsymbol{\beta}_{\mathbf{1j}} + \boldsymbol{u}_{\boldsymbol{1ij}}$$

$$\boldsymbol{\pi}_{\mathbf{2ij}} = \boldsymbol{\beta}_{\mathbf{2j}} + \boldsymbol{u}_{\boldsymbol{2ij}}$$

where $\boldsymbol{\beta}_{\mathbf{0j}}$ is the average intercept coefficient across all the students within the j-th kindergarten classroom; $\boldsymbol{\beta}_{\mathbf{1j}}$ is the average piece1 regression coefficient across all the students within the j-th kindergarten classroom, and $\boldsymbol{\beta}_{\mathbf{2j}}$ is the average piece2 regression coefficient across all the students within the j-th kindergarten classroom.

We further obtained the corresponding average coefficient estimates across all kindergarten classrooms, as presented[1].

Level 3 (classroom level)

$$\boldsymbol{\beta}_{\mathbf{0j}} = \boldsymbol{\gamma}_{\mathbf{00}} + \boldsymbol{\gamma}_{\mathbf{01}}\mathbf{treatment}_{\boldsymbol{j}} + \boldsymbol{v}_{\boldsymbol{0j}} \quad (3)$$

$$\boldsymbol{\beta}_{\mathbf{1j}} = \boldsymbol{\gamma}_{\mathbf{10}} + \boldsymbol{\gamma}_{\mathbf{11}}\mathbf{treatment}_{\boldsymbol{j}}$$

$$\boldsymbol{\beta}_{\mathbf{2j}} = \boldsymbol{\gamma}_{\mathbf{20}} + \boldsymbol{\gamma}_{\mathbf{21}}\mathbf{treatment}_{\boldsymbol{j}}$$

where $\boldsymbol{\gamma}_{\mathbf{00}}$, $\boldsymbol{\gamma}_{\mathbf{10}}$, and $\boldsymbol{\gamma}_{\mathbf{20}}$ are the average intercept, piece1 and piece2 coefficients across all kindergarten classrooms assuming a nonsignificant treatment effect.

As stated previously, one of the main purposes of the Project ELLA was to examine the effectiveness of the enhanced practice setting (i.e., the treatment condition) on EWPV. To examine this treatment effect, we included the treatment variable in the level-3 equations, given that the randomization was at the classroom/school level. In other words, students from the same kindergarten classroom received the exact same treatment or control materials. $\mathbf{Treatment}_{\boldsymbol{j}}$ is a dummy-coded variable with treatment condition coded as 1 and control condition coded as 0. Hence, if there is a significant treatment effect at intercept, we expect that $\boldsymbol{\gamma}_{\mathbf{01}}$ will not be zero and the intercept for the control condition will be $\boldsymbol{\gamma}_{00}$whereas the intercept for the treatment condition will be ($\boldsymbol{\gamma}_{00}$+ $\boldsymbol{\gamma}_{\mathbf{01}}$). Similarly, if there are significant treatment effects at both piece1 and piece2, we would expect that both $\boldsymbol{\gamma}_{11}$and $\boldsymbol{\gamma}_{21}$will not be zero and the average piece1 coefficient will be $\boldsymbol{\gamma}_{\mathbf{10}}$ for the control condition and $\boldsymbol{\gamma}_{\mathbf{10}} + \boldsymbol{\gamma}_{\mathbf{11}}$ for the treatment condition, the same as the average piece2 coefficient with $\boldsymbol{\gamma}_{\mathbf{20}}$ for the control condition and $\boldsymbol{\gamma}_{\mathbf{20}} + \boldsymbol{\gamma}_{\mathbf{21}}$ for the treatment condition.

By substituting Equations (2) and (3) back into equation (1), we can get the following overall ***average*** (or ***mean***) model:

$$\begin{aligned}\hat{\mathrm{E}}\mathbf{WPV}_{\boldsymbol{tij}} = {} & \boldsymbol{\gamma}_{\mathbf{00}} + \gamma_{\mathbf{01}}\mathbf{treatment}_{j} + \boldsymbol{\gamma}_{\mathbf{10}}\ \mathbf{piece1}_{\boldsymbol{tij}} + \\ & \boldsymbol{\gamma}_{\mathbf{11}}\mathbf{treatment}_{\boldsymbol{j}}{}^{*}\ \mathbf{piece1}_{\boldsymbol{tij}} \\ & + \boldsymbol{\gamma}_{\mathbf{20}}\ \mathbf{piece2}_{\boldsymbol{tij}}\ +\ \boldsymbol{\gamma}_{\mathbf{21}}\mathbf{treatment}_{j}{}^{*}\ \mathbf{piece2}_{\boldsymbol{tij}} \quad (4)\end{aligned}$$

The corresponding random effect variances that capture the variation at different levels are as follows:

[1] The reason of including only one random effect (i.e., $\boldsymbol{v}_{\boldsymbol{0j}}$) at the classroom level in equation (3) is to have a simpler model (in terms of the number of random effects) to avoid the potential convergence issue due to the large number of variance and covariance estimates of the random effects. Additionally, according to our experience, the variance estimates of the higher level non-intercept random effects are generally very small and non-significant and trying to estimate these tiny (and possibly non-significant) random effect variances will likely lead to non-converged result.

$V(e_{tij}) = \sigma^2$ (within-student-level variance with the identity structure assumption)

$V(u_{0ij}) = \tau_{00}$ (between-student-level intercept variance)

$V(u_{1ij}) = \tau_{11}$ (between-student-level piece1 variance)

$V(u_{2ij}) = \tau_{22}$ (between-student-level piece2 variance)

$V(v_{0ij}) = \theta^2$ (kindergarten classroom-level variance). We used the R package xxM (Mehta, 2013) to analyze our data. (The corresponding output for the model may be found in Appendix A1).

Results of Model 1

As presented in **Table 2** in the Model 1 (3-Lv HLM) column, almost all the regression coefficients were significant (with the 95% confidence interval [CI], not including zero) except $\boldsymbol{\gamma_{01}}$ (i.e., the treatment effect at the beginning of kindergarten). Hence, the overall average piecewise model for the control group (i.e., $\textbf{treatment}_j = 0$) was:

$$\hat{\mathbf{E}}\mathbf{WPV}_{tij} = \mathbf{435.6} + \mathbf{13.75\ piece1}_{tij} + \mathbf{9.64\ piece2}_{tij}$$

whereas the overall average piecewise model for the treatment group (i.e., $\textbf{treatment}_j = 1$) was:

$$\hat{\mathbf{E}}\mathbf{WPV}_{tij} = \mathbf{435.6 - 2.43(1)} + \mathbf{13.75\ piece1}_{tij} + \mathbf{2.41(1)} \\ {}^{*}\ \mathbf{piece1}_{tij} + \mathbf{9.64\ piece2}_{tij} + \mathbf{1.60(1)}\ {}^{*}\ \mathbf{piece2}_{tij},$$

which could be further reduced to:

$$\hat{\mathbf{E}}\mathbf{WPV}_{tij} = \mathbf{433.17 + 16.16\ piece1}_{tij} + \mathbf{11.24\ piece2}_{tij}.$$

Based on the average models as presented above and in Table 2, we have learned that the average EWPV for the control group at the beginning of kindergarten was 435.6 whereas the average EWPV score for the treatment group was slightly (but not significantly) lower (2.43 points lower). We have also learned that the average growth rate (or change) in EWPV was not a linear trend given that the regression coefficients of the two pieces were quite different from each other for both treatment and control groups (i.e., $\mathbf{13.75\ piece1}_{tij} + \mathbf{9.64\ piece2}_{tij}$ for the control condition and $\mathbf{16.16\ piece1}_{tij} + \mathbf{11.24\ piece2}_{tij}$ for the treatment condition). That is, we found a faster growth or improvement rate of EWPV within the kindergarten grade year and a slower growth rate of EWPV after kindergarten (i.e., from first to third grade) for both conditions, except that the students in the treatment condition, on average, showed greater improvement at the end of the kindergarten (16.16 points for the treatment condition vs. 13.75 points for the control condition) as well as at the end of first to third grade (11.24 points for the treatment condition vs. 9.64 points for the control condition). These differences in growth rates show the effectiveness of the Project ELLA enhanced

TABLE 2 | Summary of 3-Level HLM, CCREM, and xxM-UN1 model results.

	Model 1: 3-Lv HLM		**Model 2: CCREM**		**Model 3: xxM-UN1**	
FIXED						
Intercept (γ_{00})	435.60*	[432.91, 438.28]	436.99*	[434.31, 439.67]	437.07*	[434.16, 440.00]
Piece 1 (γ_{10})	13.75*	[11.96, 15.54]	13.15*	[11.36, 14.94]	13.12*	[11.51, 14.72]
Piece 2 (γ_{20})	9.64*	[8.95, 10.34]	9.47*	[8.23, 10.71]	9.66*	[8.90, 10.44]
Treatment (γ_{01})	−2.43	[−7.41, 2.61]	−3.12	[−7.20, 1.03]	−7.06*	[−11.96, -1.94]
P1 × Treat (γ_{11})	2.41*	[0.01, 4.80]	3.42*	[1.04, 5.81]	3.46*	[1.32, 5.63]
P2 × Treat (γ_{21})	1.60*	[0.59, 2.62]	0.59	[−1.36, 2.53]	1.42*	[0.23, 2.57]
RANDOM						
Student						
Intercept (τ_{00})	137.36		157.04		193.44	
P1 (τ_{11})	2.13		3.52		29.37	
Cov(Int, P1)	−17.10		−23.50		−39.80	
P2 (τ_{22})	0.08		0.16		3.66	
Cov(Int, P2)	−3.22		−5.02		−13.56	
Cov(P1, P2)	0.40		0.75		8.27	
Class (θ^2/ψ^2)	97.39		64.49		–	
K	–		–		185.37	
Grade 1	–		–		12.32	
Grade 2	–		–		10.17	
Grade 3	–		–		5.63	
Within (σ^2)	165.12		145.92		–	
MODEL FIT						
Deviance	25,959		25,889		25,423	
AIC	25,987		25,917		25,479	
BIC	26,071		26,002		25,648	

Lv: 3 level. Confidence intervals were obtained using profile likelihood method in xxM.

materials and practice on improving the students' EWPV over time.

In general, researchers are more interested in the significance of the mean part (i.e., the regression coefficients) and pay less attention to the variance part of the model. Nevertheless, the variance part carries as much important information as the mean part (e.g., treatment effect is sometimes found in the variance part instead of the mean part of the model (Hedeker and Mermelstein, 2007), and the misspecification of the variance, in part, may lead to a biased estimation not only of the fixed effects (i.e., the regression coefficients) but also of the random effect variances (Sivo et al., 2005), which may further affect the significance tests of the regression coefficients (Kwok et al., 2007).

Given that we analyzed the data as a three-level, strictly hierarchical model, the corresponding variance estimates for the different levels are presented in **Table 2** under the 3-Lv HLM column: $\sigma^2 = 165.12$ (within-student-level variance with the identity structure assumption), $\tau_{00} = 137.36$ (between-student-level intercept variance), $\tau_{11} = 2.13$ (between-student-level piece1 variance), $\tau_{22} =.08$ (between-student-level piece2 variance), and $\theta^2 = 97.39$ (***kindergarten*** classroom-level variance).

All these variances were statistically significant, which indicates a significant amount of variation within students across all the repeated measures and between students across all kindergarten classrooms. Consistent with many previous longitudinal studies using multilevel models, we found that the intercept variance (i.e., $\tau_{00} = 137.36$) was, in general, substantial larger than the variances of the two growth pieces (i.e., $\tau_{11} = 2.13$ and $\tau_{22} = 0.08$).

There are a couple of limitations to this model. First, it only partially takes into account the classroom effect (i.e., only kindergarten), which may lead to biased estimation of both regression coefficients and the random effect variances. Moreover, only modeling the kindergarten effect restricts the possibility of modeling the other grade-level effects, such as the potential carryover effect from previous grade levels (e.g., first grade) to later EWPV score (e.g., measured at third grade).

MODEL 2: THE CROSS-CLASSIFIED RANDOM EFFECT MODEL (CCREM)

Another way to analyze this longitudinal data set is to apply the cross-classified random effect model (CCREM; Luo and Kwok, 2012). Although CCREM has been proposed for many years, this model is still not commonly applied in educational studies. In our study, we also provided useful information on how this model can be and was applied to a real, large scale randomized controlled longitudinal dataset. Unlike Model 1, which assumes a strictly hierarchical structure with repeated EWPV measures nested within students who further nested only within their kindergarten classrooms, the CCREM takes into account the classroom effects over time as a whole by creating a classroom crossed factor. In other words, instead of only considering the kindergarten classroom effect, the CCREM considers all (from kindergarten to third grade) classroom effects and assumes that at a given time point the only classroom effect present is the one at that particular time point. Additionally, classroom effects at different time points are interchangeable and, therefore, form one source of random effect variance. The setup of this model is similar to that of Model 1, as illustrated below.

Level 1 (repeated-measure level)

$$\begin{aligned}\mathbf{EWPV}_{t(ij)} &= \boldsymbol{\pi}_{\mathbf{0(ij)}} + \boldsymbol{\pi}_{\mathbf{1(ij)}}\,\mathbf{piece1}_{t(ij)} + \boldsymbol{\pi}_{\mathbf{2(ij)}}\,\mathbf{piece2}_{t(ij)}\\ &\quad + \mathbf{e}_{t(ij)},\end{aligned} \tag{5}$$

where EWPV is the target outcome variable for the t-th repeated measure from the i-th student of the j-th classroom and piece1 and piece2 are the time variables with the exact same coding scheme. The major difference between this model and Model 1 is the presentation and meaning of the subscript (specifically the "j" subscript). Unlike in Model 1 where the j subscript is only for a particular kindergarten classroom, the j subscript in Model 2 represents a particular classroom of any grade level (i.e., from kindergarten to third grade). That is, the students are no longer nested only within the kindergarten classrooms, as shown in **Figure 1B**. Instead, as shown in **Figure 1A**, the repeated measures are now nested within the i-th students and the j-th classroom whereas student and classroom are now crossed with each other. Hence the subscripts i and j in Equation (5) are now grouped in the parentheses (ij). For example, Student S1 in **Figure 1A** has three repeated measures (O_{11}, O_{12} and O_{13}), as does Student S2 (O_{21}, O_{22} and O_{23}). Students S1 and S2 are in different kindergarten classrooms (KC_1 for S1 and KC_2 for S2) but are in the same classroom in first grade ($G1C_1$) and are assigned to different classrooms second grade ($G2C_1$ for S1 and $G2C_2$ for S2). Hence, the repeated measures (i.e., Os) are nested both within students (S1 and S2) and classrooms (KC_1, KC_2, $G1C_1$, $G2C_1$ and $G2C_2$), whereas students and classrooms are crossed instead of nested.

Given that student and classrooms are crossed with each other, the level-2 model in CCREM includes both students and classrooms simultaneously as presented below:

Level 2 (student and classroom level)

$$\begin{aligned}\boldsymbol{\pi}_{0(\mathrm{ij})} &= \boldsymbol{\gamma}_{\mathbf{00}} + \boldsymbol{\gamma}_{\mathbf{01}}\mathbf{treatment}_{\boldsymbol{j}} + \boldsymbol{u}_{\boldsymbol{0i}} + \boldsymbol{v}_{\boldsymbol{0j}}\\ \boldsymbol{\pi}_{\mathbf{1(ij)}} &= \boldsymbol{\gamma}_{\mathbf{10}} + \boldsymbol{\gamma}_{\mathbf{11}}\mathbf{treatment}_{\boldsymbol{j}} + \boldsymbol{u}_{\boldsymbol{1i}}\\ \boldsymbol{\pi}_{\mathbf{2(ij)}} &= \boldsymbol{\gamma}_{\mathbf{20}} + \boldsymbol{\gamma}_{\mathbf{21}}\mathbf{treatment}_{\mathbf{j}} + \boldsymbol{u}_{\boldsymbol{2i}},\end{aligned} \tag{6}$$

where $\boldsymbol{\gamma}_{\mathbf{00}}$, $\boldsymbol{\gamma}_{\mathbf{10}}$, and $\boldsymbol{\gamma}_{\mathbf{20}}$ are the average intercept, piece1 and piece2 coefficients across all classrooms, assuming the non-significant treatment effect. On the other hand, given that the randomization was at the classroom level, we included the dummy-coded treatment variable, treatment$_j$, in the level-2 equations. Hence, if there is a significant treatment effect at intercept, $\boldsymbol{\gamma}_{00}$will be the intercept for the control condition whereas $\boldsymbol{\gamma}_{00}+\ \boldsymbol{\gamma}_{\mathbf{01}}$ will be the intercept for the treatment condition. Similarly, if there are significant treatment effects at both piece1 and piece2, the average piece1 coefficient will be $\boldsymbol{\gamma}_{\mathbf{10}}$ for the control condition and $\boldsymbol{\gamma}_{\mathbf{10}} + \boldsymbol{\gamma}_{\mathbf{11}}$ for the treatment condition; the same holds for the average piece2 coefficient, with $\boldsymbol{\gamma}_{\mathbf{20}}$ for the control condition and $\boldsymbol{\gamma}_{\mathbf{20}} + \boldsymbol{\gamma}_{\mathbf{21}}$ for the treatment condition.

By substituting Equation (6) back into Equation (5), we obtained the following overall ***average*** (or ***mean***) model, which is almost the same as Equation (4) under Model 1:

$$\begin{aligned}\hat{\mathbf{E}}\mathbf{WPV}_{t(ij)} \;=\; & \boldsymbol{\gamma}_{\mathbf{00}} + \boldsymbol{\gamma}_{\mathbf{01}}\mathbf{treatment}_{j} + \boldsymbol{\gamma}_{\mathbf{10}}\ \mathbf{piece1}_{t(ij)} \\ & + \boldsymbol{\gamma}_{\mathbf{20}}\ \mathbf{piece2}_{t(ij)} \;+\; \boldsymbol{\gamma}_{\mathbf{11}}\mathbf{treatment}_{j} * \mathbf{piece1}_{t(ij)} \\ & + \boldsymbol{\gamma}_{\mathbf{21}}\mathbf{treatment}_{j} * \mathbf{piece2}_{t(ij)} \end{aligned} \tag{7}$$

The corresponding random effect variances are as follows:

$V(\boldsymbol{e}_{t(ij)}) = \sigma^2$ (within-student-level variance with the identity structure assumption)

$V(\boldsymbol{u}_{0i}) = \tau_{00}$ (between-student-level intercept variance)

$V(\boldsymbol{u}_{1i}) = \tau_{11}$ (between-student-level piece1 variance)

$V(\boldsymbol{u}_{2i}) = \tau_{22}$ (between-student-level piece2 variance)

$V(\mathbf{v}_{0j}) = \psi^2$ (between-classroom-level variance).

The major difference between this CCREM model and Model 1 is with regard to the random effect part; specifically, the classroom effect $\boldsymbol{v}_{0j}$ with the corresponding variance equal to ψ^2. Even though it seems like only a slight change in the combined equation (from $\boldsymbol{v}_{0j}$ of the kindergarten random effects in Model 1 to $\boldsymbol{v}_{0j}$ of all classroom random effects in Model 2), the actual implication and the parameter estimates of Model 2 can be very different from those of Model 1 due to the variance redistribution mechanism (Luo and Kwok, 2009). The corresponding output for this model may be found in Appendix A2. Below, we highlight these differences.

Results of Model 2

The results are presented in **Table 2** in the Model 2 (CCREM) column. Instead of explaining each parameter estimate, we have highlighted the major differences between Models 1 and 2. First, the $\mathbf{treatment}_j * \mathbf{piece2}_{t(ij)}$ interaction effect is no longer significant in Model 2 ($\boldsymbol{\gamma}_{21}$ =.64 with the 95% CI covered zero) compared with Model 1. This nonsignificant interaction effect indicates that the rate of change or improvement in the EWPV was the same for both treatment and control groups after kindergarten.

In addition to the regression coefficient, some of the estimates of the random effect variances were quite different between the two models: Model 2 had a larger intercept variance ($\tau_{00} = 157.04$ compared with Model 1 $\tau_{00} = 137.36$), a smaller classroom variance ($\psi^2 = 64.49$ compared with Model 1 $\theta^2 = 97.39$), and a smaller within-student variance ($\sigma^2 = 145.92$ compared with Model 1 $\sigma^2 = 165.12$). These differences in the variance estimates between the two models are likely the result of the variance redistribution mechanism (Luo and Kwok, 2009). Although the number of parameters are the same in the two models, the meaning and setup (in terms of the design matrix) of the random effects, especially the classroom random effects, can result in quite different variance estimates which, in turn, can lead to different standard error estimates and tests of significance of the regression coefficients.

Regarding the limitation of this model, unlike Model 1 which only takes into account the kindergarten classroom effect, Model 2 is able to fully take the classroom effect into account. However, it does assume an acute classroom effect (i.e., it will not carry over in later grades). In other words, once a student changes grade (i.e., classroom), he/she will get a new classroom effect. The classroom effect at kindergarten is independent of the classroom effect at grade 1, for example. Also, all classroom effects regardless the grade (or time) have exactly the same variance given that they are treated as a whole or a single crossed factor, even though conceptually the classrooms at different grades/times may have different effects on the EWPV scores.

Ideally, we wanted to analyze this data set with four classroom crossed factors but, in reality, the specification for this model is not straightforward, especially when using the common MLM packages. Moreover, the model estimates only the variance for the classroom factors, not the other effects, such as the potential carryover effect from the previous classrooms on later EWPV scores.

MODEL 3: xxM-UN1 PIECEWISE LATENT GROWTH

Whereas the nesting relationship holds in cross-sectional data, in longitudinal settings the relationship between students and classrooms is not pure. To make things more complicated, students' scores at a given time point, say second grade, are not only influenced by the classroom effect at second grade, but also potentially by the classroom effects at both kindergarten and first grade. Furthermore, the effect of the classroom may diminish, such that the impact of first grade may have a stronger effect on the second-grade scores than at third grade. Such a model would include five crossed random effects (i.e., one at the student level and four at the classroom level, including kindergarten, first-, second-, and third-grade random effects) and would need to allow the classroom effects to vary across time. None of the default models from the standard statistical packages can fully capture the key feature of this model.

Similar to Model 1, Model 3 (also see **Figure 2**) has also effectively captured the growth pattern and the treatment by pieces interaction effects after taking into account the data dependency. However, both Models 1 and 2 may not be the most optimal approach to analyze these data given some of the restricted assumptions. For example, they both assume that the residuals have a constant variance and are independent across time (i.e., an identity structure for the within-student variance-covariance structure). Moreover, they assume a constant classroom effect across time without any impact or carryover effect (from one grade level to the next).

The first limitation can be addressed by specifying a different residual covariance structure than the default one (see Kwok et al., 2007), which can be done in most multilevel software programs, such as HLM, SAS, and SPSS, as well as with the latent growth models under the SEM framework. The second limitation requires specification of multiple, crossed random effects to capture the potential non-constant classroom effects, which cannot be easily estimated in standard multilevel software[2]. Nevertheless, recent developments in the *n*-level SEM and the

[2] To our best knowledge, across all the commercial SEM related software, only M*plus* and Stata (the "gsem" routine) can handle cross-classified data with limited number of crossed factors (e.g., Mplus can only handle two crossed factors).

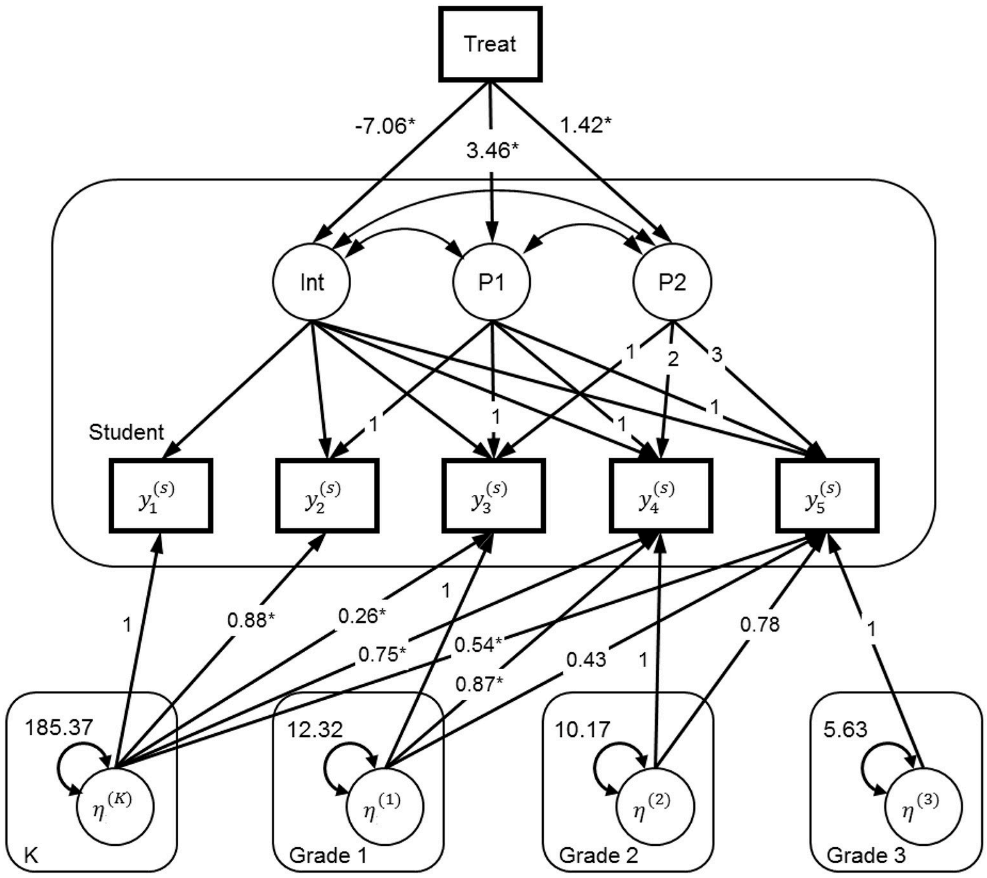


FIGURE 2 | Path diagram for the model accommodating carryover classroom effects with five levels. y_1, beginning of kindergarten; y_2, end of kindergarten; y_3, end of first grade; y_4, end of second grade; y_5, end of third grade. The rounded-corner boxes: Student, Student level; K, Kindergarten classroom level; Grade 1, Grade 1 classroom level; Grade 2, Grade 2 classroom level; Grade 3 classroom level.

corresponding R package xxM (Mehta, 2013) have provided the potential to specify more complex multilevel models, including Model 3, as presented in **Figure 2**.

The model specification in xxM requires a combination of multilevel and SEM conventions. Due to its complexity, we only discuss the portions that are relevant to our model. First, it requires the longitudinal data to be in the wide rather than the long format (Kwok et al., 2008) in order to model complex residual covariance structures. This is identical to the latent growth modeling approach using SEM. Second, it requires a separate data set at each level, which is similar to the setup in HLM. In our model, we want to model five levels: the student level and four classrooms levels, including kindergarten (class-K), first grade (class-G1), second grade (class-G2), and third grade (class-G3). Third, it requires model specification at each level, and also for each pairwise combination of levels. For example, for a latent growth model with an additional classroom-level random effect, we have $\mathbf{y}_i^{(1)} = \boldsymbol{\Lambda}^{(1,1)}\boldsymbol{\eta}_i^{(1)} + \boldsymbol{\Lambda}^{(1,2)}\boldsymbol{\eta}_i^{(2)} + \boldsymbol{\varepsilon}_i^{(1)}$, where $\mathbf{y}_i^{(1)}$ is the vector of the outcome scores of student i from Time 1 to Time 5, $\boldsymbol{\Lambda}^{(1,1)}$ is a fixed pattern matrix for our piecewise growth model, $\boldsymbol{\eta}_i^{(1)}$ is the vector of latent growth factor scores (i.e., intercept, piece1, and piece2) with mean $\boldsymbol{\alpha}^{(1)}$ and variance-covariance matrix $\boldsymbol{\Psi}^{(1,1)}$, and $\boldsymbol{\varepsilon}_i^{(1)}$ is the student-level error terms. The superscripts $^{(1)}$ and $^{(1,1)}$ denote a student-level model. At the classroom-level there is one latent variable $\eta_i^{(2)}$ denoting the random intercept, with mean $\boldsymbol{\alpha}^{(2)} = 0$ and variance $\boldsymbol{\psi}^{(2)}$, and with direct paths on $\mathbf{y}_i^{(1)}$ through the between-class-K-student-level matrix $\boldsymbol{\Lambda}^{(1,2)} = [1,\ 1,\ 1,\ 1,\ 1]^{\mathrm{T}}$.

Because of the complexity associated with using xxM, we skip the model equations here to focus more on the conceptual formulation instead. The R code for fitting the model is presented in Appendix B.

Student-Level Model

At the student level, we have a piecewise latent growth model for the five EWPV measurement occasions, which is equivalent to the piecewise growth model with random intercept and random coefficients for both piece1 and piece2, as opposed to lme4 (Bates et al., 2015), which requires the residual covariance structure to be a constant σ^2 over time (i.e., an identity (ID) structure) as follows:

$$\sigma^2 \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

In xxM, we can model many other kinds of structure, such as freely estimating the residual variances for different time points (i.e., the first-order unstructured [UN1] structure), as presented here in which the residual variances vary across time measures.

$$\begin{bmatrix} \sigma_A^2 & 0 & 0 & 0 & 0 \\ 0 & \sigma_B^2 & 0 & 0 & 0 \\ 0 & 0 & \sigma_C^2 & 0 & 0 \\ 0 & 0 & 0 & \sigma_D^2 & 0 \\ 0 & 0 & 0 & 0 & \sigma_E^2 \end{bmatrix}$$

This seems to be a more realistic choice than the ID structure. The treatment condition that was assigned at kindergarten or as a class-K level variable predicts the intercept and the two piecewise growth factors (P1 and P2 in **Figure 2**). The corresponding path coefficients (of the paths/arrows from Treat to the growth latent factors in Figure 2) are conceptually equivalent to $\boldsymbol{\gamma_{01}}$ (**treatment**$_j$), $\boldsymbol{\gamma}_{11}$(**treatment**$_j$ * **piece1**$_{t(ij)}$) and $\boldsymbol{\gamma}_{21}$(**treatment**$_j$ * **piece2**$_{t(ij)}$) in the previous two models.

Four Classroom-Level Models

At the class-K level, we have a random intercept factor $\eta^{(K)}$ that accounts for the variance at all five time points due to clustering at kindergarten. We let the effect of such clustering differ across time points, which is achieved by allowing the direct paths (or factor loadings) from $\eta^{(K)}$ to be different on the five measurement occasions. It is reasonable to expect that the effect will diminish across time, which means that the factor loadings should be decreasing. At the class-G1 level, we again have a random intercept factor $\eta^{(G1)}$ that accounts for the clustering at first grade. Because classroom effect at first grade cannot affect prior performance (i.e., at kindergarten), the factor loadings from $\eta_1^{(G1)}$ to the first two measures are fixed at zero. Similar procedures are carried out for the remaining two random intercept factors, $\eta^{(G2)}$ and $\eta^{(G3)}$, as shown in **Figure 2**.

Results of Model 3

Given that the interpretation of the coefficients of the average or mean model is exactly the same as in the previous two models, we will focus more on the differences between Model 3 and the other two models. First, as shown in **Table 2**, all the fixed effects or regression coefficients were statistically significant with the 95% profile likelihood CI not covering zero. Specifically, when comparing the fixed effect estimates of Model 3 with those of the other two models, both coefficients of treatment ($\boldsymbol{\gamma_{01}} = -7.06$) and Piece2 × Treatment ($\boldsymbol{\gamma_{21}} = 1.42$) became significant.

Figure 3 contains the estimated average models for both groups based on the estimates from the xxM-UN1 column in **Table 2**. As shown in **Figure 3**, the treatment group (the dashed line group) has lower EWPV scores at the beginning of kindergarten, and the growth (or improvement) rate of this group is faster than the control group at both pieces (i.e., the kindergarten piece and the first- to second-grade piece). The difference between the two groups on EWPV diminished as time passed, and by the end of second grade, the two lines crossed, which indicated no differences between the two groups. In other words, even though the treatment students started with significantly lower EWPV scores at the beginning of kindergarten, they caught up with their control group counterparts (by the end of second grade) and might even outperform them at the later time points. Notice also that the width of the CI is smaller for terms involving piece2 (compared with the corresponding terms involving piece1). This is likely a result of the decreasing classroom effect across time.

Another major difference between Model 3 and the previous two models is found in the variance part of the model: not only does Model 3 contain more random effects (i.e., four different classroom effects for the four different grades), but the sizes of the variance estimates (i.e., τ_{00}, τ_{11}, and τ_{22}) are quite different from Models 1 and 2. As shown in **Table 2**, unlike the other two models with a single classroom variance, Model 3 contained four classroom variances for the four different grades, respectively.

A closer analysis of these classroom variances reveals that the kindergarten variance was the largest whereas the third-grade variance was the smallest. This trend and the substantial differences across grades may partly be the result of missing data—the missing data rate increased as time passed, and with fewer students at the later time points or grades, it is not surprising to see the diminished variance estimates. Other potential reasons may include the developmental process (i.e., students learn more when they grow older) and plausible treatment effect (e.g., students become more homogeneous/similar to each other when they respond to the treatment materials). Further investigation of this issue is needed.

For the same random effect variances (i.e., τ_{00}, τ_{11}, and τ_{22}), Model 3 had substantially larger estimates than the other two models. This again may be the result of the variance redistribution mechanism (Luo and Kwok, 2009) due to the additional classroom variances. Given that the standard errors (SEs) of the fixed effect estimates (or regression coefficients) are a function of the random effect variances, the additional significant coefficients (i.e., $\boldsymbol{\gamma_{01}}$ & $\boldsymbol{\gamma_{21}}$) in Model 3 are likely the results of these different variance estimates, which can directly affect the tests of significance of these coefficients.

In addition to the fixed and random effect estimates commonly found in the traditional multilevel models and presented in **Table 2**, we further examined the potential carryover effect using xxM due to its flexibility of specifying more complex multilevel models. As shown in **Figure 2**, the direct paths (arrows) from each classroom factor to the individual time measures can be viewed as examining the carryover effect; that is, the effect from the previous grade classroom to the current and later time EWPV scores. For model identification, we constrained the direct path of the current time measure to 1.0 (e.g., fixing the

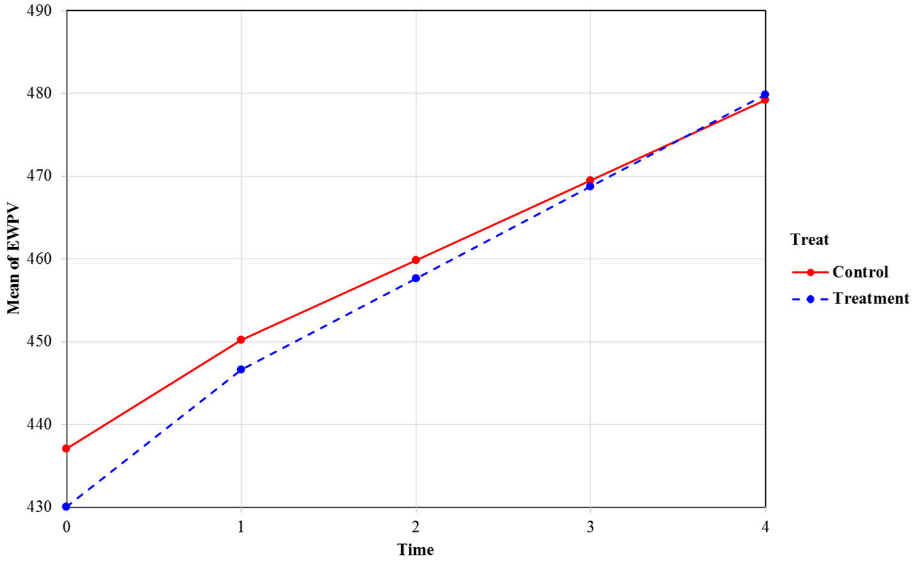


FIGURE 3 | Mean trajectories of EWPV scores by the two treatment conditions. Value labels for the time axis: 0, beginning of kindergarten; 1, end of kindergarten; 2, end of first grade; 3, end of second grade; 4, end of third grade.

kindergarten effect to Y1 [K-begin] to 1.0, while freely estimating other paths). As shown in **Figure 2**, the freely estimated direct paths from kindergarten to all the time measures (K-end, 1st-end, 2nd-end, and 3rd-end) were significant, with the largest effect at the immediate post measure (i.e., the end of kindergarten EWPV score) followed by weaker effects at later time measures.

We found a similar pattern for the first-grade factor (i.e., larger direct path coefficient to the immediate post measure followed by smaller coefficient to later time measures), even though the direct path coefficients were not all significant, possibly as a result of the smaller sample sizes at this grade and the later grade levels. Similar non-significant direct effects were also found for the second-grade factor.

These significant and non-significant carryover effects at different grade levels had some important and practical implications. For example, the many significant carryover effects from kindergarten may reflect the importance of the timing (i.e., the start of the intervention) and the potential longitudinal effect of the intervention. In other words, we may not see the same treatment effect if the intervention starts at another grade level as opposed to the beginning of kindergarten. Moreover, the significant paths from kindergarten to later-grade EWPV scores may reveal the importance of the kindergarten classroom experience, which may relate to ELL students' reading performance in the later grades, and further examination of this will be needed.

We compared the three models by using information criteria; namely, the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). Certain guidelines apply to interpreting the absolute difference of the information criteria (i.e., ΔIC) between two competing models. For example, Burnham and Anderson (1998) suggested that when ΔAIC between two compared models is larger than 4, we can establish that the model with smaller AIC is better than the other model with larger AIC. Likewise, Raftery (1996) pointed out that the ΔBIC between two competing models should be at least 2 to indicate a real difference. Based on these guidelines, we found that Model 3 fit the data the best given the smallest AIC and BIC values across all three models.

DISCUSSION

In this study, we first described the complexity of the educational data, especially in longitudinal settings, which can result in data with a non-strictly hierarchical but more complex multilevel structure. With the use of the ELLA data, we demonstrated the importance of capturing the complex data structure by examining three different models with different random effect specification.

As stated, researchers are generally interested in the overall average model (or the mean part of the model containing the regression coefficients), but they fail to pay close attention to the variance part of the model. Yet, the variance part also carries important information, such as the implication of the developmental process. We have discussed and shown the importance of carefully specifying the random part of the model, which could affect estimation of the random effect variances and further affect estimates of the standard errors of the regression coefficients and the corresponding significance tests of these coefficients. For example, we found that both Models 1 and 3 had significant treatment by pieces interaction effects whereas Model 2 only had significant treatment by piece1 interaction effect and only contained some but not all significant coefficients. This finding provides evidence that only partially addressing the complex data structure may result in lower statistical power and loss of some important findings such as the treatment by growth piece (i.e. piece2 covering changes from the end of first to end of third grade) interaction effect.

Another advantage of modeling the classroom effect by grade levels separately (i.e., Model 3) instead of as a whole (e.g., Model 2 using CCREM) is that it allows researchers to investigate interesting phenomena that cannot be captured by the mean part of the model. For example, the decreasing classroom or grade variances over time may reflect the important developmental process. For example, the high heterogeneity (or variation) among students at the beginning of kindergarten may be the result of the diverse backgrounds and experiences the students have before they entered formal schooling. Once they are exposed to the formal grade-school curriculum in addition to their natural cognitive development, the variation among the students may become smaller, which in turn, may lead to a reduction in grade-level variances over time.

This is a plausible explanation, but further systematic investigation on the change in the variances is needed to validate this interpretation. Again, researchers should not only focus on the mean part of the model (i.e., the significance of the regression coefficients), but also, they should examine different random effect structure, which may provide different perspectives and even lead to new research questions for the target phenomena.

Moreover, we have shown how to incorporate the carryover effect in the model via the xxM program. The pattern of the carryover effect has shed light on some important and practical design issues, such as the timing of the study and the potential longitudinal impact of the intervention. For example, the only significant carryover effects from the kindergarten factor to the later time measures may suggest the importance of starting this type of intervention at kindergarten (rather than at other/later grade level). In fact, such carryover impact was also supported by empirical evidence on Project ELLA students' subsequent learning as they matriculated to grade 5 (e.g., Tong et al., 2014).

Despite the important results presented here, there are a few limitations to the study. First, even though xxM is a very powerful software for very complex multilevel data, its lack of model-fit indices (e.g., RMSEA and CFI) restricts researchers to evaluate their models only based on the deviance statistic and the information criteria. Similarly, an appropriate standardized effect size measure for this type of complex data structure has not yet been developed. Another major limitation is that we only used real data for the demonstration. Thus, the actual impact of various factors such as the magnitude of the data dependency (or intra-class correlation) and the missing data rate over time can only be further examined by thoughtfully planned simulation studies. Moreover, the carry-over effects found in Model 3 (also see **Figure 2**) are in arbitrary metric, and researchers need to be cautious when interpreting these findings. Besides xxM, a similar type of model (Model 3) may possibly specify and analyze with non-SEM Bayesian based programs such as STAN (Carpenter et al., 2017). Further investigation on whether and how effective this alternative approach on fitting the same type of carry-over effect model to similar real, large scale randomized controlled longitudinal data will be needed.

When analyzing complex longitudinal data, especially those from different educational settings, researchers generally focus only on the mean part (i.e., the regression coefficients) while ignoring the equally important random part (i.e., the random effect variances) of the model. Throughout this paper, we have addressed the importance of adequately taking the complex data structure into account by carefully specifying the random part of the model—not only can it affect the variance estimates, the standard errors, and the tests of significance of the regression coefficients, it can also offer additional information such as the potential developmental process and the carryover effect. We used xxM, which allowed us to estimate different grade level variances (i.e., from kindergarten to third grade, separately) and the potential carryover effect from each grade factor to the later time measures of the EWPV scores. In closing, we encourage researchers to look beyond the mean part of the model (i.e., the regression coefficients) and explore the variance part of the model that may lead them to different perspectives or even new information of the phenomena they are studying.

AUTHOR CONTRIBUTIONS

O-MK and ML are the lead authors who wrote most of the manuscript and conducted all the analyses. Other coauthors contributed on providing the data and related information FT, RL-A, and BI) and offering constructive feedback to the manuscript FT, RL-A, BI, MY, and Y-CY).

FUNDING

The data described in this article is supported by Grant R305P030032, funded by the Institute of Education Sciences (IES). The contents, findings, and opinions expressed in this article are those of the authors and do not necessarily represent the positions or policies of IES. We thank our IES project coordinator and grant project coordinators and the graduate assistants, teachers, paraprofessionals, parents, and school and district officials who made this research possible. We are grateful for the insightful suggestions provided by reviewers who read earlier versions of this article.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2018.00790/full#supplementary-material

REFERENCES

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01

Bollen, K. A., and Curran, P. J. (2006). *Latent Curve Models: A Structural Equation Modeling Perspective*. Hoboken, NJ: Wiley.

Bryk, A. S., and Raudenbush, S. W. (1992). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Newbury Park, CA: Sage.

Burnham, K. P., and Anderson, D. R. (1998). *Model Selection and Inference: A Practical Information-Theoretic Approach.* New York, NY: Springer-Verlag.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., et al. (2017). Stan: A probabilistic programming language. *J. Stat. Softw.* 76, 1–32. doi: 10.18637/jss.v076.i01

Curran, P. J. (2003). Have multilevel models been structural equation models all along? *Multiv. Behav. Res.* 38, 529–569. doi: 10.1207/s15327906mbr3804_5

Duncan, T. E., Duncan, S. C., and Strycker, L. A. (2006). *An Introduction to Latent Variable Growth Curve Modeling: Concepts, Issues, and Applications, 2nd Edn*. Mahwah, NJ: Lawrence Erlbaum.

Hedeker, D., and Mermelstein, R. J. (2007). "Mixed-effects regression models with heterogeneous variance: analyzing ecological momentary assessment data of smoking," in *Modeling Contextual Effects in Longitudinal Studies*, eds T. D. Little, J. A. Bovaird, and N. A. Card (Mahwah, NJ: Lawrence Erlbaum), 183–206.

Kwok, O., Luo, W., and West, S. G. (2010). Using modification indexes to detect turning points in longitudinal data: A Monte Carlo study. *Struc. Equat. Model.* 17, 216–240. doi: 10.1080/10705511003659359

Kwok, O., Underhill, A. T., Berry, J. W., Luo, W., Elliott, T. R., and Yoon, M. (2008). Analyzing longitudinal data with multilevel models: an example with individuals living with lower extremity intra-articular fractures. *Rehabil. Psychol.* 53, 370–386. doi: 10.1037/a0012765

Kwok, O., West, S. G., and Green, S. B. (2007). The impact of misspecifying the within-subject covariance structure in multilevel longitudinal multilevel models: a Monte Carlo study. *Multiv. Behav. Res.* 42, 557–592. doi: 10.1080/00273170701540537

Lara-Alecio, R. (2003). *English Language and Literacy Acquisition (Project ELLA).* Washington, DC: U.S. Department of Education. Available online at: http://epsy.tamu.edu/sites/epsy.tamu.edu/files/Lara-Alecio-Project%20ELLA.pdf

Luo, W., and Kwok, O. (2009). The impacts of ignoring a crossed factor in analyzing cross-classified data. *Multivariate Behav. Res.* 44, 182–212. doi: 10.1080/00273170902794214

Luo, W., and Kwok, O. (2012). The consequences of ignoring individuals' mobility in multilevel growth models: A Monte Carlo study. *J. Educ. Behav. Stat.* 37, 31–56 doi: 10.3102/1076998610394366

Mehta, P. (2013). *xxM: Structural Equation Modeling for Dependent Data* (R package version 0.6.0) [Computer program]. Available online at: http://xxm.times.uh.edu/

Meyers, J., and Beretvas, S. N. (2006). The impact of inappropriate modeling of cross-classified data structures. *Multiv. Behav. Res.* 41, 473–497. doi: 10.1207/s15327906mbr4104_3

Preacher, K., Wichman, A., MacCallum, R., and Briggs, N. (2008). *Latent Growth Curve Modeling.* Thousand Oaks, CA: Sage.

Raftery, A. E. (1996). Bayesian model selection in social research. *Sociol. Methodol.* 25, 111–163. doi: 10.2307/271063

Sayer, A. G., and Willett, J. B. (1998). A cross-domain model for growth in adolescent alcohol expectancies. *Multiv. Behav. Res.* 33, 509–543. doi: 10.1207/s15327906mbr3304_4

Sivo, S., Fan, X., and Witta, L. (2005). The biasing effects of unmodeled ARMA time series processes on latent growth curve model estimates. *Struct. Equ. Model.* 12, 215–231. doi: 10.1207/s15328007sem1202_2

Snijders, T. A. B., and Bosker, R. (1999). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. Newbury Park, CA: Sage.

Texas Education Code (1995). 74th Leg., Ch. 260, Section 29.056, § 1

Tong, F., Irby, B. J., Lara-Alecio, R., and Koch, J. (2014). A longitudinal study of integrating literacy and science for fifth grade Hispanic current and former English language learners: from learning to read to reading to learn. *J. Educ. Res.* 107, 410–426. doi: 10.1080/00220671.2013.833072

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 16 May 2017
doi: 10.3389/fpsyg.2017.00735

The Development and Validation of the Online Shopping Addiction Scale

Haiyan Zhao[1,2], Wei Tian[3] and Tao Xin[3]*

[1] Faculty of Psychology, Beijing Normal University, Beijing, China, [2] Beijing Education Examinations Authority, Beijing, China, [3] Collaborative Innovation Center of Assessment toward Basic Education Quality at Beijing Normal University, Beijing, China

We report the development and validation of a scale to measure online shopping addiction. Inspired by previous theories and research on behavioral addiction, the Griffiths's widely accepted six-factor component model was referred to and an 18-item scale was constructed, with each component measured by three items. The results of exploratory factor analysis, based on Sample 1 (999 college students) and confirmatory factor analysis, based on Sample 2 (854 college students) showed the Griffiths's substantive six-factor structure underlay the online shopping addiction scale. Cronbach's alpha suggested that the resulting scale was highly reliable. Concurrent validity, based on Sample 3 (328 college students), was also satisfactory as indicated by correlations between the scale and measures of similar constructs. Finally, self-perceived online shopping addiction can be predicted to a relatively high degree. The present 18-item scale is a solid theory-based instrument to empirically measure online shopping addiction and can be used for understanding the phenomena among young adults.

Keywords: online shopping addiction, behavioral addiction, internet addiction, compulsive buying, scale development

OPEN ACCESS

Edited by:
Oi-Man Kwok,
Texas A&M University, USA

Reviewed by:
Dylan Molenaar,
University of Amsterdam, Netherlands
Donald Sharpe,
University of Regina, Canada

****Correspondence:***
Tao Xin
xintao@bnu.edu.cn

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *12 November 2016*
Accepted: *24 April 2017*
Published: *16 May 2017*

Citation:
Zhao H, Tian W and Xin T (2017) The Development and Validation of the Online Shopping Addiction Scale. Front. Psychol. 8:735. doi: 10.3389/fpsyg.2017.00735

INTRODUCTION

Initial definitions of addiction focused on drug ingestion or intake of substances (Walker, 1989; Rachlin, 1990). Some behaviors of this kind can be regarded as substance addiction. Other behaviors that do not involve drug ingestion also have the potential for addiction, albeit with psychological and physiological correlates similar with drug ingestion (Shaffer et al., 2004). Research on the non-substance-related or behavioral addiction is growing. Examples of such addiction include game playing (e.g., Fisher, 1994; Lemmens et al., 2009), gambling (Griffiths, 1995; Brand et al., 2005), overeating (Orford, 2001), exercise (Adams and Kirkby, 2002; Berczik et al., 2012), internet use (e.g., Young, 1998; Beard, 2005), shopping (Clark and Calleja, 2008; Davenport et al., 2012), cellphone use (Rutland et al., 2007; Chóliz, 2010), and work (Andreassen et al., 2010; Andreassen, 2014).

With the popularity of the wired lifestyle (Bellman et al., 1999), online shopping addiction (OSA) has begun to appear as a new behavioral addiction. According to Rose and Dhandayudham (2014), OSA may have negative influences not only on an individual's daily life and social life, but also on their economic status. Consequently, the diagnosis, intervention, and treatment of OSA are of great importance. Thus, a reliable and valid instrument to measure OSA is essential. To operationalize OSA, it is helpful to consider some similar constructs, such as internet addiction (IA) and compulsive buying (CB).

Over the last 15 years, there is debate about whether IA is a genuine addiction (Griffiths and Pontes, 2014). Since "Gambling Disorder" has been re-classified as a disorder of addiction instead of impulse control in the latest edition of the Diagnostic and Statistical Manual of Mental Disorders

(DSM-5) (American Psychiatric Association, 2013), that re-classification suggests considering IA as a genuine addiction. According to Davis (2001), IA can be classified into specific and generalized types depending on the target of the behavior. The former IA uses the internet for particular purposes, such as online gaming, gambling, social networking, etc., while the latter IA had no specific aims.

As far as OSA is concerned, many researchers hold that OSA can be classified into the category of specific IA (Brand et al., 2014; Griffiths and Szabo, 2014; Laconi et al., 2015; Montag et al., 2015; Pontes et al., 2015). Griffiths (2000) argued that it is important to distinguish between addictions on the internet and addictions to the internet. Specifically, many people spending excessive time on the internet are not addicted to the medium itself, but use the medium to actualize other addictions (Pontes et al., 2015). From this perspective, OSA should be a specific type of IA.

CB refers to a tendency toward long-term, repeated buying behavior, which has become the individual's primary response to negative events and emotions (O'Guinn and Faber, 1989; Black, 2007; Müller et al., 2015; Trotzke et al., 2015). Many researchers regard CB as a behavioral addiction (Demetrovics and Griffiths, 2012; Lo and Harvey, 2012; Starcke et al., 2013; Rose and Dhandayudham, 2014), while others emphasize that a typical behavioral addiction involves much time spent in thinking about engaging in the behavior and is therefore characterized by intense preoccupation (Sussman et al., 2010). Although, OSA and CB are rather similar in both external manifestation and internal features, there may still be subtle differences between them, such that OSA may be confined to the internet, while CB has no such restriction.

Assessment of Online Shopping Addiction

To the best of our knowledge, no specialized instruments for OSA exist, although some relevant instruments need to be mentioned. The first is the Bergen Shopping Addiction Scale (Andreassen et al., 2016), which was designed to measure the core criterion and components of shopping addiction. The scale consists of 28 items, four for each of the seven addiction criteria listed in the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV-TR; American Psychiatric Association, 2000). The content of the items reflect contemporary shopping habits and the scale show good validity and reliability. Other relevant instruments include those used by researchers examining OSA as a specific IA. For instance, to assess OSA, the corresponding subscale of the Shorter PROMIS questionnaire (Christo et al., 2003) was modified by adding the terms "internet" or "online." The resulting ten-item scale proved to be reasonably reliable (Laconi et al., 2015). In another instance, Montag et al. (2015) used the short version of the Gaming Addiction Scale (Lemmens et al., 2009) as a blueprint and constructed several specific IA scales including OSA by exchanging the word "game" in each item with each specific form of IA. The resulting seven-item scale also had high consistency across different samples.

All the instruments mentioned above followed the common practice in behavioral addiction, in which instruments were constructed based on certain factor models, such as the six-factor model (Brown, 1993; Griffiths, 1996) or the seven-addiction criteria of DSM-IV (American Psychiatric Association, 2000). Although, the authors of these instruments claim or imply the existence of a specific IA, they employ a similar construct structure to the generalized IA. The particularity of instruments is item wordings, whether a specific or generalized IA they are designed for. The particularity of specific IA was never examined or resolved in the content or expressions of items, let alone the relationship and distinction between specific and generalized IA. Nevertheless, there are instruments constructed based on different structures, such as the Facebook Addiction Scale (Torsheim et al., 2012) and the Game Addiction Scale (Lemmens et al., 2009). The former was based on a one-factor solution and the latter was on a second-order structure. Both these scales exhibited sound reliability and validity, which implied that other structures might also be plausible for these specific IA.

The aim of the present study was to construct a specialized instrument for OSA. We sampled college students as participants, because in China, individuals of this age are independent of their parents and they are skillful in using the internet. Previous research also suggests that the similar disorder, CB, usually has an onset in one's 20s and turns into a chronic disorder in their later years (Black, 2007).

Development of the Online Shopping Addiction Scale

In the present study, we view OSA as a specific type of IA. Referring to the definitions of other behavioral addiction (Griffiths, 2005), we defined OSA as a tendency of excessive, compulsive and problematic shopping behavior via the internet that results in consequences associated with economic, social, and emotional problems. Addictive shoppers still fail to control their excessive online shopping behaviors despite problematic consequences.

During the development of the OSA scale, we employed the six-factor model of behavioral addiction (Brown, 1993; Griffiths, 1996, 2002). This model holds that the following six elements are necessary for operational definitions of addictive behaviors: salience, mood modification, tolerance, withdrawal, conflict, and relapse. Salience means that addiction behaviors have become the most important activity in addicts' lives, preoccupying their thoughts, dominating their cravings, and demonstrating an excessive occurrence. Mood modification refers to the subjective experience of conducting addictive behaviors, e.g., feeling high, buzz or exciting, quiet, released, numb or even depressed after fulfillment. Tolerance means that in order to achieve the effects equal to that in the past, addicts have to increase the amount of the activities. Withdrawal indicates the unpleasant sensation and/or physiological reaction after the addictive behaviors are cut off or restricted. Conflict indicates the interpersonal conflict between individuals and others together with the intrapsychic conflict within individuals. Relapse refers to the tendency of returning to the original behavioral modes after dropping or restricting addictive behaviors; the addiction will burst into the most severe degree after relapse (Brown, 1993; Griffiths, 1996, 2005).

In practice, we began by developing a pool of possible items, reflecting these six factors. When constructing the preliminary items, we also referred to several behavioral addiction or compulsive buying instruments (e.g., Christo et al., 2003; Lemmens et al., 2009; Torsheim et al., 2012; Müller et al., 2015; Andreassen et al., 2016). We considered the circumstances of online shopping and adapted the impacts of other behavioral addiction to the context of online shopping. Three linguistic specialists and three researchers specializing in psychometrics reviewed the raw items. Then we modified the items elaborately based on their comments and data from pre-tests. In all, the review and modification process took three rounds. The preliminary scale contained 27 items, four or five for each of the six components. Each item included a sentence about the strength, effect, and internal or external influence of online shopping. The final scale was listed in **Table 1** together with the component that each item belonged to.

METHODS

Participants and Sampling

We collected three groups of participants from about 30 colleges in China for the purpose of scale development and validation. Sample 1, the exploratory sample, consisted of 999 students (744 female); their age varied between 18 and 28 (Mean = 21.05; *SD* = 1.87). Sample 2, the confirmatory sample, consisted of 854 students (575 female); their age ranged from 18 to 28 (Mean = 21.36; *SD* =2.04). Sample 3, the validation sample, consisted of 328 students (159 female); their age ranged from 18 to 28 (Mean = 21.79; *SD* = 2.25).

Measures

Demographic Questionnaire

A demographic questionnaire collected information about demographic variables, which also included an item for rating self-perceived degree of OSA.

TABLE 1 | Mean scores, standard deviation, measures of distribution, and the corrected item-total correlation for the18-item online shopping addiction scale based on the exploratory sample.

Subscales	Item	Item content	M	*SD*	Skewness	Kurtosis	CITC
Salience	S1	When I am not shopping online, I keep thinking about it	3.44	1.10	−0.61	−0.42	0.45
	S2	I frequently think about how to spare more time or money to spend in online shopping	3.27	1.17	−0.23	−0.84	0.53
	S3	Online shopping is important for my life	3.80	1.03	−0.89	0.30	0.38
Tolerance	T1	Recently, I have an urge to do more and more online shopping	2.38	1.18	0.47	−0.86	0.55
	T2	I spend more and more time in online shopping	2.13	1.13	0.75	−0.43	0.56
	T3	Recently I often shop online unplanned	2.62	1.29	0.10	−1.32	0.53
Mood modification	M1	When I feel bad, online shopping can make me feel good	3.28	1.11	−0.33	−0.61	0.54
	M2	When I am feeling down, anxious, helpless or uneasy, I shop online in order to make myself feel better	2.23	1.25	0.60	−0.94	0.45
	M3	Online shopping can help me to temporarily forget the troubles in real life	2.32	1.22	0.48	−0.98	0.57
Withdrawal	W1	When I can't do online shopping for certain excuses, I will get depressed or lost	2.19	1.14	0.67	−0.58	0.63
	W2	Life without online shopping for some time would be boring and joyless for me	2.23	1.22	0.66	−0.74	0.68
	W3	I will feel restless or depressed when attempting to shop online but unable to achieve	2.46	1.22	0.34	−1.11	0.55
Relapse	R1	I have tried to cut back or stop my online shopping, but failed	2.16	1.09	0.77	−0.22	0.59
	R2	I have decided to do online shopping less frequently, but not managed to do so	2.06	1.03	0.77	−0.26	0.59
	R3	If I cut down the amount of online shopping in one period, and then start again, I always end up shopping as often as I did before	1.86	1.04	1.09	0.32	0.67
Conflict	C1	My productivity for work or study has decreased as a direct result of online shopping	1.68	0.88	1.35	1.57	0.45
	C2	I have once quarreled with my parents for my online shopping	1.42	0.83	2.31	5.26	0.25
	C3	I have cut off my time with parents and friends for my online shopping	1.53	0.79	1.69	2.81	0.52

M, Mean; SD, Standard Deviation; CITC, Corrected Item-Total Correlation.

The Online Shopping Addiction Scale

The preliminary version of the scale included 27 items, each rated on a five-point Likert scale (1 = completely disagree, 2 = disagree, 3 = neither disagree nor agree, 4 = agree, and 5 = completely agree). As can be seen in the result section, the final version consisted of 18 items. Internal consistency (Cronbach's alpha) was 0.90 and 0.95 for the confirmatory and the validation samples, respectively.

The Compulsive Buying Scale

The Edward's Compulsive Buying Scale (Ridgway et al., 2008) is a 13-item instrument assessing compulsive buying behavior. Each item was rated on a four-point scale with anchor of 1, with higher scores indicating a tendency toward compulsive buying. Cronbach's alpha was 0.93 for the validation sample.

The Internet Addiction Test

This test consists of 20 items, each rated on a five-point Likert scale. High scores on the test indicated serious problems caused by the internet (Young, 1998). Cronbach's alpha was 0.96 for the validation sample.

Procedure

All three samples were collected online. We distributed the survey link in different students groups at about 30 colleges in China during the winter of 2016 and the spring of 2017. The study was carried out in accordance with the Helsinki Convention and the Norwegian Health Research Act. The protocol and the survey packet were reviewed and approved by the Ethics Committee of the research team's university. The survey could be accessed online for 1 week for each sample. The purpose of the study was displayed on the top of the first webpage of the e-questionnaire. Participants were deemed to be consent to participate if and only if the survey was completed. All questions were collected anonymously and no money or other incentives were given. For both the exploratory and confirmatory samples, the demographic questionnaire and the preliminary OSA scale were posted, while for the validation sample, the Compulsive Buying Scale and the Internet Addiction Test were additionally included.

Statistics

With sample 1, item discrimination based on classical test theory (CTT) was used to select the most effective items to form a highly reliable and valid instrument. The selected items were then used to explore the possible factor structure with the exploratory factor analysis (EFA). With sample 2, the factor structure explored was justified with the confirmatory factor analysis (CFA). In addition, the psychometric properties and validity evidence of the OSA scale were also assessed via sample 3.

Model fit was evaluated in terms of goodness of fit statistics, specifically the chi-square, comparative fit index (CFI), the Tucker-Lewis index (TLI), root mean square error of approximation (RMSEA), and the standardized root mean square residual (SRMR). Statistical analyses were conducted with Mplus 7.2 (Muthén and Muthén, 2013). The criteria for good fit statistics were non-significant chi-square, CFI ≥ 0.96, TLI ≥ 0. 95, and RMSEA ≤ 0.06 (Hu and Bentler, 1999), and for acceptable fit were CFI ≥ 0.90, TLI ≥ 0 .90, and RMSEA ≤ 0.08 (Vandenberg and Lance, 2000; Marsh et al., 2004). Additionally, SRMR values below 0.08 were typically considered to reflect reasonable model fit (Hu and Bentler, 1999).

RESULTS

Scale Construction

Item Analysis

The preliminary scale included 27 items developed with both empirical and theoretical underpinnings. Item assessment was intended to identify items that would be problematic to remain in the following analyses. We considered items with low corrected item-total correlations as problematic and excluded them from subsequent analyses. In addition, we also reviewed the item endorsement frequencies (noting those items whose frequencies fell below 90% or above 10%) to detect whether there was adequate item variance across participants and skewed responses.

As a result, we identified 18 satisfactory items based on Sample 1. Three items were included within each of the six sub-domains to assure content validity. **Table 1** showed that the corrected item-total correlation (CITC) of the 18 items ranged from 0.25 to 0.68. Unfortunately, responses were found to depart somewhat from normal distributions, with skewness levels ranging from −0.89 to 2.31.

Exploratory Factor Analysis

Based on Sample 1, the final 18 items were used to conduct EFA. Considering some of the 18 items had non-normal distribution, EFA was run using the robust weighted least square mean and variance (WLSMV) estimation. Item responses were treated as categorical variables, and polychoric correlations were analyzed. CFI, TLI, RMSEA, and SRMR were reported for each factor solution in **Table 2**. Oblique rotations using the GEOMIN method were generated because the intended OSA factors were correlated.

Among all the seven solutions EFA extracted, both the six-factor and the seven-factor structure underlies the newly developed OSA scale judging by the criteria of model fit index. Furthermore, the seven-factor is somewhat superior to the six-factor solution from a model comparison perspective. On the one hand, however, including an additional 7th factor in the structure only contributes 3.7% percent of more variance to be explained. On the other hand, from the perspective of substantive theory, the factor loadings pattern for the current six-factor solution shown in **Table 3** was nearly the ideal simple theoretical factor structure. Therefore, we finally chose the six-factor structure as the optimal factor structure. In particular, the explored factors could be approximately defined as salience, withdrawal, relapse, conflict, tolerance, and mood modification. As to the six-factor solution, the corresponding eigenvalues for sample correlation matrix were 7.73, 1.81, 1.01, 0.91, 0.83, 0.77, respectively.

Although, the nearly perfect simple six-factor structure was explored, it should be noted that there were still some obvious cross-loading associated with a few items, such as item T2, T3, M1, W3, R1, R2, and C1. For example, item T2 does not only load on tolerance, but also loads high on conflict. Item T3 loads

TABLE 2 | Summary of model fit information for exploratory factor analysis.

	Chi-Square	df	CFI	TLI	RMSEA	90% RMSEA	SRMR	Chi-Square compared
1-factor	2390.70*	135	0.84	0.82	0.13	[0.125, 0.134]	0.09	
2-factor	952.09*	118	0.94	0.93	0.08	[0.079, 0.089]	0.05	846.92*(1-factor against 2-factor)
3-factor	686.01*	102	0.96	0.94	0.08	[0.070, 0.081]	0.04	245.97*(2-factor against 3-factor)
4-factor	441.35*	87	0.98	0.96	0.06	[0.058, 0.070]	0.03	212.72*(3-factor against 4-factor)
5-factor	281.50*	73	0.99	0.97	0.05	[0.047, 0.060]	0.02	139.47*(4-factor against 5-factor)
6-factor	194.59*	60	0.99	0.98	0.05	[0.040, 0.055]	0.02	83.66*(5-factor against 6-factor)
7-factor	121.26*	48	1.00	0.99	0.04	[0.030, 0.048]	0.01	68.91*(6-factor against 7-factor)

*CFI, comparative fit index; TLI, Tucker-Lewis index; RMSEA, Root Mean Square Error of Approximation; SRMR, Standardized Root Mean Square Residual; *Significant at 5% level.*

TABLE 3 | Exploratory factor analysis factor loadings for the six-factor model of the online shopping addiction scale using weighted least square mean and variance with GEOMIN method rotation.

Items	f_1	f_2	f_3	f_4	f_5	f_6
S1	0.40	0.02	0.15	0.05	−0.01	0.17
S2	0.76	−0.01	0.17	0.09	−0.05	0.00
S3	0.88	0.02	−0.08	−0.05	0.02	0.01
T1	−0.01	0.78	0.03	0.28	0.00	−0.01
T2	0.17	0.27	−0.01	−0.02	0.11	0.45
T3	0.14	0.11	0.07	−0.04	0.46	0.08
M1	0.37	−0.02	0.60	0.01	0.08	−0.19
M2	−0.02	0.04	0.59	−0.21	0.26	0.03
M3	0.05	0.03	0.60	0.03	−0.04	0.22
W1	0.07	−0.02	0.27	0.63	0.03	0.04
W2	0.21	0.16	0.24	0.30	0.15	0.05
W3	−0.05	0.11	0.33	0.50	−0.03	0.04
R1	0.08	−0.09	−0.01	0.42	0.56	−0.08
R2	−0.07	0.04	−0.02	0.39	0.53	0.05
R3	−0.02	0.02	0.07	0.06	0.72	0.15
C1	0.01	0.05	−0.10	0.30	0.05	0.51
C2	−0.07	−0.13	0.07	0.07	0.03	0.53
C3	0.01	−0.01	0.12	0.00	0.02	0.76

only weak on tolerance, but high on relapse. Similarly, item M1 loads significantly on its theoretical dimension and on salience dimension, and the same pattern occurred to W3, R1, R2, and C1. Additionally, the covariance matrix for the EFA Sample was attached in Appendix.

Confirmatory Factor Analysis

Based on the results of EFA and the substantial theory, the six-factor structure was replicated through CFA using weighted least square mean and variance (WLSMV) estimation. For this six factor model in sample 2, the Chi-Square test was 825.22 with degrees of freedom as 120, the CFI was 0.95, the TLI was 0.94, and the RMSEA was 0.08, suggesting acceptable model fit. The corresponding standardized factor loadings of the six-factor model were showed in **Figure 1**. As can be seen, the factor loadings were high and ranged from 0.55 to 0.84. At the same time, the intercorrelations between six factors was also presented in **Table 4**, which showed that the six factors were highly correlated. Additionally, the covariance matrix for the CFA Sample was attached in Appendix.

Psychometric Properties of the Online Shopping Addiction Scale

Internal Consistency

The Cronbach's alpha was 0.95 for samples 3, which indicated a high degree of internal consistency. **Table 5** showed that alpha varied between 0.71 and 0.84 for different subscales, indicating that internal consistencies of the most subscales were satisfactory. It should be noted that Cronbach alpha was 0.71 for "Mood modification" subscale and was 0.76 for "Salience" subscale. In view that each subscale consists only three items, these coefficients were acceptable. Inter-correlations between subscales ranged from 0.48 to 0.78, all statistically significant at the 0.01 level.

Concurrent Validity

We assessed Concurrent validity against scores on the Edward's Compulsive Buying Scale (Ridgway et al., 2008) and the Internet Addiction Test (Young, 1998), as external measures of constructs similar to OSA. The evaluation of concurrent validity relies on an understanding of how strongly constructs should or should not relate to each other. **Table 6** showed the correlations between the total score of 18-item OSA scale, its subscale scores and the other two scales. Correlations with the CB scale were higher than those with the IA test, which indicated the construct of OSA actually focused not only on excessive shopping behaviors generally, but also fulfill this inclination on the internet.

Predictive Validity

The self-perceived online shopping addiction was an important indicator to assess group differences and practical prediction. Here the item asked the participants to describe their self-perceived degree of OSA in 1 = severe, 2 = moderate, 3 = mild, and 4 = no addiction.

We first assessed the utility of the OSA scale by examining group differences between four responses with variance analysis technique. The homogeneity of group variance was first examined with Levene's test. As was showed in **Table 7**, the Levene's test was significant and suggested that the group variances were not equal. However, the variance ratio computed with largest variance divided by the smallest one was 2.83,

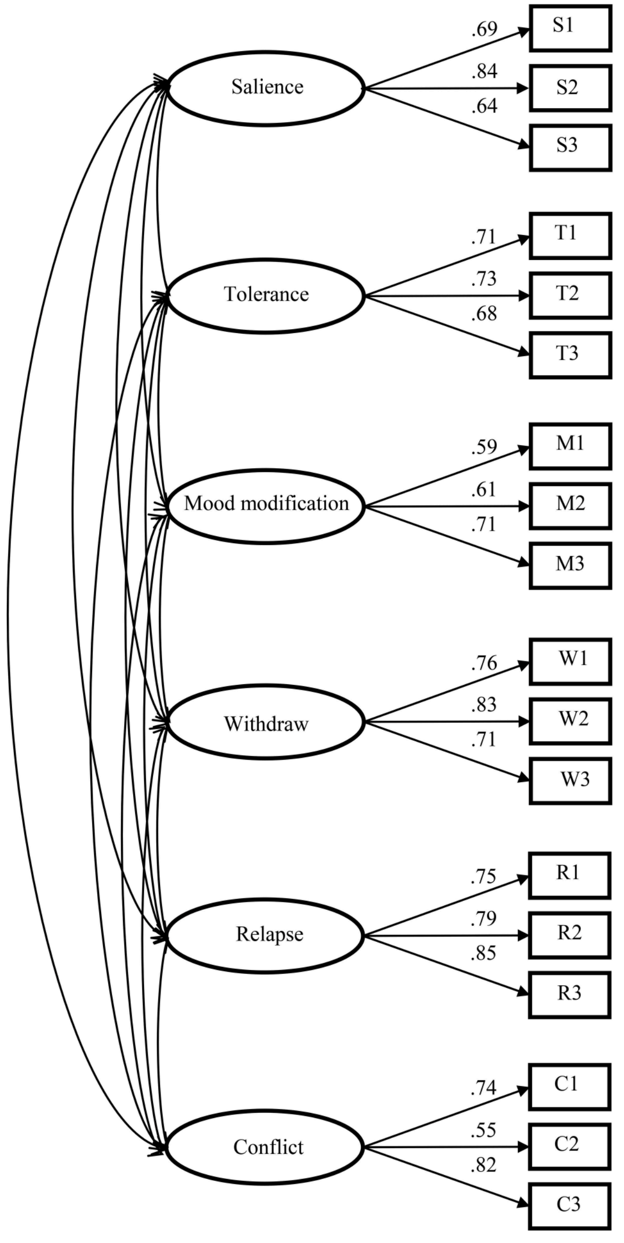


FIGURE 1 | The six-factor model and its standardized factor loadings.

TABLE 4 | The intercorrelations between six factors based on the confirmatory factor analysis.

Factors	f_1	f_2	f_3	f_4	f_5	f_6
f_1	1					
f_2	0.67	1				
f_3	0.77	0.85	1			
f_4	0.68	0.88	0.84	1		
f_5	0.58	0.92	0.76	0.85	1	
f_6	0.39	0.83	0.68	0.73	0.82	1

TABLE 5 | Internal consistencies (Cronbach's alpha) and the inter-correlations for subscales based on the validity Sample.

Subscales	Alpha	1	2	3	4	5	6
1 Salience	0.76	1					
2 Tolerance	0.84	0.71	1				
3 Mood modification	0.71	0.67	0.70	1			
4 Withdrawal	0.83	0.68	0.77	0.70	1		
5 Relapse	0.84	0.64	0.77	0.65	0.78	1	
6 Conflict	0.83	0.48	0.67	0.59	0.69	0.71	1

All correlations were computed with composite scores and were significant at 0.01 level.

TABLE 6 | Correlations between the total score and sub-scale scores of the online shopping addiction scale and the scores of the compulsive buying scale and the internet addiction test.

	Compulsive buying	Internet addiction
Salience	0.54	0.51
Tolerance	0.69	0.58
Mood modification	0.66	0.55
Withdrawal	0.66	0.57
Relapse	0.67	0.57
Conflict	0.64	0.53
OSA	0.75	0.64

All correlations were significant at 0.01 level.

TABLE 7 | Descriptive statistics for the self-perceived OSA and the test of homogeneity of variance.

Self-perceived	*M*	*N*	*SD*	Levene's statistic	df1	df2	Sig.
1	66.91	11	18.99	3.73	3	324	0.01
2	51.74	78	13.28				
3	42.40	134	12.55				
4	30.11	105	11.29				

which was too small to worry much about. The group difference was significant at 0.01 level with $F_{(3,\,324)} = 60.92$. Finally, all possible pair-wise comparisons of means was conducted using the Least Significant Difference (LSD) test. The *Post-hoc* comparisons showed that the means were ordered as expected.

We further tested the agreement between self-perceived and model-predicted membership via the classification table from logistic regression. We recoded severe and moderate self-ratings as addiction, and the mild and no addiction ratings as non-addiction. The subsequent logistic regression of self-perceived addiction on the item performances yielded the percentage of correctly prediction shown in **Table 8**. The overall correctly predicted percentage was 79.60, implying the precision of the scale for screening and diagnosis.

TABLE 8 | Consistency of Self-Diagnostic and Predicted Addiction by the OSA Scale.

		Predicted addiction		Percentage correct
		0	**1**	
Self-perceived addiction	0	220	19	92.10
	1	48	41	46.10
Overall percentage				79.60

DISCUSSION

We conducted the present study to develop a reliable and valid instrument for OSA. Based on previous research on behavioral addiction, we adopted the widely accepted six-factor component model (Brown, 1993; Griffiths, 1996) and constructed an 18-item OSA scale, with each component measured by three items. The results of EFA indicated that the six-factor structure underlay the newly developed scale from perspectives of model comparison and substantive theory. Moreover, the results of CFA also demonstrated that the six-factor structure fit the data well. In terms of reliability and validity, the Cronbach's alpha suggested that the scale was highly reliable and the concurrent validity was also satisfactory as indicated by correlations between the scale and measures of similar constructs. Finally, the OSA scale scores predicted the self-perceived online shopping addiction to a relative high degree. We conclude that the present 18-item scale is a solid theory-based instrument to empirically measure online shopping addiction.

It has been argued that since the late 1990s that most people who spend excessive time on the internet are not addicted to the medium itself, but use it to fulfill specific addiction, such as video game playing or shopping (Griffiths, 1999, 2000). In order to clarify the structure of OSA, it was necessary to make clear the relationship between generalized and specific IA. The present study held that although specific type had their special objects of thinking, feeling, and activities, they shared common components with generalized type. Consequently, the target of the present study, online shopping addiction, indeed could be represented by the six-factor component model.

The results showed the commonalities between specific internet addiction and generalized internet addiction in that both constructs had the common six components of behavioral addition. However, what is the particularity of online shopping addiction as a type of specific internet addiction? When we examined the content of the items of two validity scales, it was obvious that items of the OSA scale shared more similarities with those of the Internet Addiction Test. However, the analysis of concurrent validity showed that online shopping addition was more relevant to compulsive buying than to internet addiction. This contrast suggested that the similarities between OSA and IA were more superficial than those between OSA and CB. Furthermore, OSA is more than a form of internet addiction. In nature, it is a form of shopping addiction and addicts use the internet mainly to fulfill their problematic shopping inclination. This hinted us that in terms of the diagnosis and intervention of specific internet addiction, more emphasis should be put on its peculiarities.

When we examined the agreement between self-perceived and model-predicted membership, the misclassified ratio was a little high for the self-reported addiction category. This can be partly due to the nature of OSA and the characteristic of the present sample. As a behavioral addiction, the base rate of OSA in non-clinical sample could be rather low, just like the similar disorder, compulsive buying (Black, 2007). Furthermore, it was inevitable for some participants answered the survey in a casual way, especially for an online version. This was possible since a large percent rated themselves as having severe or moderate addiction, which indicated that some participants had the inclination to exaggerate their own status. Therefore, self-perceived degree of OSA could only be a very gross indicator and it finally contributed to the relative high ratio for misclassification.

Limitations and Suggestions for Future Research

Based on the six-factor component model for behavioral addiction (Brown, 1993; Griffiths, 1996), we constructed a specialized instrument for online shopping addiction and acquired reasonable results. However, the appropriateness of the structure was unable to cover up a couple of deficiencies in the present study. Firstly, although the present scale asked the examinees to respond on a five-point scale, there were still substantial amounts of participants choosing to respond on only four categories or less. It is still unclear why participants choose to response in less categories than demanded. This may result from that it is difficult for participants to distinguish the subtleness of adjacent category when the number of categories amounts to some extent. It may also result from that some participants are just inclined to response in very limited categories, which is a common style under the circumstances of Chinese culture. In the next phase, we plan to change the scale into a few formats to explore the optimal number of response category.

In the present study, the results of EFA did conform to a simple six-factor structure. The factor pattern was clear with only a few items embodying substantial cross-loadings. These cross-loadings could probably due to the nature of the construct. Since different factors were set to be oblique during the EFA process, items pertaining to highly correlating factors could be confused instinctively. Furthermore, when we examined the correlation matrix for exploratory sample thoroughly, it was clear that some item pairs turned out to correlate rather high, such as M1 and S2, C3, and T2, and T3 and R3. These high correlations between items from different factors might be due to phrasing and content of items, which could possible contribute to significant cross-loadings. Cross-loadings can contaminate the structure of the construct, which was especially true for the "tolerance" element. Among the three items of this subscale, there was one item loading rather high on "relapse" and another item loading significantly both on "withdrawal" and "tolerance." As for these cross-loadings and possible flaw in items, we also plan to do more research to recognize which loadings naturally do not adhere to the structure, and to clarify the structure of construct and the scale.

Although, some studies (Black, 2007; Clark and Calleja, 2008; Lejoyeux and Weinstein, 2010) have shown that shopping addiction and CB overlap to a great extent, we hypothesized that they were distinct, albeit related, constructs. The same principle applied to the relationship between OSA and other behavioral addictions, especially IA. The fact that OSA correlated significantly with CB or IA does not mean that there was causal relationship existing between them, all of which could be the results of more substantial and fundamental factors, such as personality traits (Sun and Wu, 2011; Rose and Dhandayudham, 2014). During the development of the present scale, the similarities between OSA and other behavioral addictions were taken into full consideration, while the peculiarities of OSA still remain to be emphasized in future research.

Additionally, all examinees of the present study were college students, which were rather similar in characteristics and backgrounds. In near future, we plan to distribute the survey in more heterogeneous examinees to acquire more generalized results. Furthermore, new IRT-based methods and techniques would also be implemented in future studies to improve the whole quality of the scale, including the exploration of the elaborated characteristics of the items, the item functioning differences across genders and the measurement invariance across groups.

AUTHOR CONTRIBUTIONS

TX led the design and implement of the study, including the literature search, analysis, interpretation of the data, drafting, writing, and revising. All authors contributed to the design (HZ, WT, and TX), questionnaires construction and elaboration (HZ, WT, and TX), data collection (HZ and WT), analysis (HZ and WT), interpretation of data (HZ, WT, and TX), and writing and revising the work critically (HZ, WT, and TX). All authors read and approved the final version of the work to be published (HZ, WT, and TX) and agreed to be accountable for all aspects of the work in ensuring that any question to the accuracy of the work is appropriately investigated and resolved (HZ, WT, and TX).

FUNDING

The manuscript is supported by National Natural Science Foundation of China (Award number 31371047).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpsyg.2017.00735/full#supplementary-material

REFERENCES

Adams, J., and Kirkby, R. J. (2002). Excessive exercise as an addiction: a review. *Addict. Res. Theory* 10, 415–438. doi: 10.1080/1606635021000032366

Andreassen, C. S. (2014). Workaholism: an overview and current status of the research. *J. Behav. Addict.* 3, 1–11. doi: 10.1556/JBA.2.2013.017

Andreassen, C. S., Griffiths, M. D., Pallesen, S., Bilder, R. M., Torsheim, T., and Aboujaoude, E. (2016). The Bergen shopping addiction scale: reliability and validity of a brief screening test. *Front. Psychol.* 6:1374. doi: 10.3389/fpsyg.2015.01374

Andreassen, C. S., and Hetland, J., and Pallesen, S. (2010). The relationship between "workaholism" basic needs satisfaction at work and personality. *Eur. J. Pers.* 24, 3–17. doi: 10.1002/per.737

American Psychiatric Association (2000). *Diagnostic and Statistical Manual of Mental Disorders, 4th Edn., Text Revision.* Washington, DC: American Psychiatric Association.

American Psychiatric Association (2013). *Diagnostic and Statistical Manual of Mental Disorders, 5th Edn*. Washington, DC: American Psychiatric Association.

Beard, K. W. (2005). Internet addiction: a review of current assessment techniques and potential assessment questions. *CyberPsychol. Behav.* 8, 7–14. doi: 10.1089/cpb.2005.8.7

Bellman, S., Lohse, G., and Johnson, E. (1999). Predictors of online buying behavior. *Commun. ACM* 42, 32–38. doi: 10.1145/322796.322805

Berczik, K., Szabó, A., Griffiths, M. D., Kurimay, T., Kun, B., Urbán, R., et al. (2012). Exercise addiction: symptoms, diagnosis, epidemiology, and etiology. *Subst. Use Misuse* 47, 403–417. doi: 10.3109/10826084.2011.639120

Black, D. W. (2007). Compulsive buying disorder: a review of the evidence. *Int. J. Neuropsychiatr. Med.* 12, 124–132. doi: 10.1017/s1092852900020630

Brand, M., Kalbe, E., Labudda, K., Fujiwara, E., Kessler, J., and Markowitsch, H. J. (2005). Decision making in patients with pathological gambling. *Psychiatry Res.* 133, 91–99. doi: 10.1016/j.psychres.2004.10.003

Brand, M., Laier, C., and Young, K. S. (2014). Internet addiction: coping styles, expectancies, and treatment implications. *Front. Psychol.* 5:1256. doi: 10.3389/fpsyg.2014.01256

Brown, R. I. F. (1993). "Some contributions of the study of gambling to the study of other addictions," in *Gambling Behavior and Problem Gambling*, eds W. R. Eadington and J. Cornelius (Reno, NV: University of Nevada Press), 241–272.

Chóliz, M. (2010). Mobile phone addiction: a point of issue. *Addiction* 105, 373–374. doi: 10.1111/j.1360-0443.2009.02854.x

Christo, G., Jones, S. L., Haylett, S., Stephenson, G. M, Lefever, R. M., and Lefever, R. (2003). The shorter PROMIS questionnaire: Further validation of a tool for simultaneous assessment of multiple addictive behaviours. *Addict. Behav.* 28, 225–248. doi: 10.1016/S0306-4603(01)00231-3

Clark, M., and Calleja, K. (2008). Shopping addiction: a preliminary investigation among maltese university students. *Addict. Res. Theory* 16, 633–649. doi: 10.1080/16066350801890050

Davenport, K., Houston, J. E., and Griffiths, M. D. (2012). Excessive eating and compulsive buying behaviours in women: an empirical pilot study examining reward sensitivity, anxiety, impulsivity, self-esteem and social desirability. *Int. J. Ment. Health Addict.* 10, 474–489. doi: 10.1007/s11469-011-9332-7

Davis, R. A. (2001). A cognitive behavioral model of pathological internet use. *Comput. Human Behav.* 17, 187–195. doi: 10.1016/S0747-5632(00)00041-8

Demetrovics, Z., and Griffiths, M. D. (2012). Behavioral addictions: past, present and future. *J. Behav. Addict.* 1, 1–2. doi: 10.1556/JBA.1.2012.1.0

Fisher, S. (1994). Identifying video game addiction in children and adolescents. *Harv. Rev. Psychiatry* 19, 545–553. doi: 10.1016/0306-4603(94)90010-8

Griffiths, M. (1999). Gambling technologies: prospects for problem gambling. *J. Gambl. Stud.* 15, 265–283.

Griffiths, M. (2002). *Gambling and Gaming Addictions in Adolescence*. Leicester: British Psychological Society/Blackwells.

Griffiths, M. D. (1995). *Adolescent Gambling*. London: Routledge.

Griffiths, M. D. (1996). Nicotine, tobacco and addiction. *Nature* 384:18. doi: 10.1038/384018a0

Griffiths, M. D. (2000). Internet addiction - Time to be taken seriously? *Addict. Res.* 8, 413–418. doi: 10.3109/16066350009005587

Griffiths, M. D. (2005). A "components" model of addiction within a biopsychosocial framework. *J. Subst. Use* 10, 191–197. doi: 10.1080/14659890500114359

Griffiths, M. D., and Pontes, H. M. (2014). Internet Addiction disorder and internet gaming disorder are not the same. *J. Addict. Res. Ther.* 5:e124. doi: 10.4172/2155-6105.1000e124

Griffiths, M. D., and Szabo, A. (2014). Is excessive online usage a function of medium or activity? *J. Behav. Addict.* 3, 74–77. doi: 10.1556/JBA.2.2013.016

Hu, L., and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struct. Equation Model.* 6, 1–55. doi: 10.1080/10705519909540118

Laconi, S., Tricard, N., and Chabrol, H. (2015). Differences between specific and generalized problematic Internet uses according to gender, age, time spent online and psychopathological symptoms. *Comput. Human Behav.* 48, 236–244. doi: 10.1016/j.chb.2015.02.006

Lejoyeux, M., and Weinstein, A. (2010). Compulsive buying. *Am. J. Drug Alcohol Abuse* 36, 248–225. doi: 10.3109/00952990.2010.493590

Lemmens, J. S., Valkenburg, P. M., and Peter, J. (2009). Development and validation a game addiction scale for adolescents. *Media Psychol.* 12, 77–95. doi: 10.1080/15213260802669458

Lo, H., and Harvey, N. (2012). Effects of shopping addiction on consumer decision- making: Web-based studies in realtime. *J. Behav. Addict.* 1, 162–170. doi: 10.1556/JBA.1.2012.006

Marsh, H. W., Hau, K.-T., and Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Struct. Equation Model.* 11, 320–341. doi: 10.1207/s15328007sem1103_2

Montag, C., Bey, K., Sha, P., Li, M., Chen, Y. F., Liu, W. Y., et al. (2015). Is it meaningful to distinguish between generalized and specific internet addiction? Evidence from a cross-cultural study from Germany, Sweden, Taiwan and China. *Asia Pac. Psychiatry* 7, 20–26. doi: 10.1111/appy.12122

Müller, A., Trotzke, P., Mitchell, J. E., de Zwaan, M., and Brand, M. (2015). The pathological buying screener: development and psychometric properties of a new screening instrument for the assessment of pathological buying symptoms. *PLoS ONE* 10:e0141094. doi: 10.1371/journal.pone.0141094

Muthén, L. K., and Muthén, B. (2013). *Mplus User's Guide, 7th Edn.* Los Angeles, CA: Muthén and Muthén.

O'Guinn, T. C., and Faber, R. J. (1989). Compulsive buying: a phenomenological exploration. *J. Consum. Res.* 16, 147–157. doi: 10.1086/209204

Orford, J. (2001). *Excessive Appetites: A Psychological View of the Addictions, 2nd Edn.* Chichester: Wiley.

Pontes, H., and Szabo, A., and Griffiths, M. D. (2015). The impact of Internet-based specific activities on the perceptions of Internet addiction, quality of life, and excessive usage: a cross-sectional study. *Addict. Behav. Rep.* 1, 19–25. doi: 10.1016/j.abrep.2015.03.002

Rachlin, H. (1990). Why do people gamble and keep gambling despite heavy losses? *Psychol. Sci.* 1, 294–297. doi: 10.1111/j.1467-9280.1990.tb00220.x

Ridgway, N. M., Kukar-Kinney, M., and Monroe, K. B. (2008). An expanded conceptualization and a new measure of compulsive buying. *J. Consum. Res.* 35, 622–639. doi: 10.1086/591108

Rose, S., and Dhandayudham, A. (2014). Towards an understanding of Internet-based problem shopping behaviour: the concept of online shopping addiction and its proposed predictors. *J. Behav. Addict.* 3, 83–89. doi: 10.1556/JBA.3.2014.003

Rutland, J. B., Sheets, T., and Young, T. (2007). Development of a scale to measure problem use of short message service: the SMS problem use diagnostic questionnaire. *CyberPsychol. Behav.* 10, 841–843. doi: 10.1089/cpb.2007.9943

Shaffer, H. J., LaPlante, D. A., LaBrie, R. A., Kidman, R. C., Donato, A. N., and Stanton, M. V. (2004). Towards a syndrome model of addiction: multiple expressions, common etiology. *Harv. Rev. Psychiatry* 12, 1–8. doi: 10.1080/10673220490905705

Starcke, K., Schlereth, B., Domass, D., Schöler, T., and Brand, M. (2013). Cue reactivity towards shopping cues in female participants. *J. Behav. Addict.* 2, 17–22. doi: 10.1556/JBA.1.2012.012

Sun, T., and Wu, G. (2011). Trait predictors of online impulsive buying tendency: a hierarchical approach. *J. Mark. Theory Pract.* 19, 337–346. doi: 10.2753/MTP1069-6679190307

Sussman, S., Lisha, N., and Griffiths, M. D. (2010). Prevalence of the addictions: a problem of the majority or the minority? *Eval. Health Prof.* 34, 3–56. doi: 10.1177/0163278710380124

Torsheim, T., Brunborg, G. S., and Pallesen, S. (2012). Development of a facebook addition scale. *Psychol. Rep.* 110, 501–517. doi: 10.2466/02.09.18.PR0.110.2.501-517

Trotzke, P., Starcke, K., Pedersen, A., Müller, A., and Brand, M. (2015). Impaired decision making under ambiguity but not under risk in individuals with pathological buying-behavioral and psychopysiological evidence. *Psychiatry Res.* 229, 551–558. doi: 10.1016/j.psychres.2015.05.043

Vandenberg, R. J., and Lance, C. E. (2000). A review and synthesis of measurement invariance literature: suggestions, practices, and recommendations for organizational research. *Organ. Res. Methods* 3, 4–70. doi: 10.1177/109442810031002

Walker, M. B. (1989). Some problems with the concept of "gambling addiction": should theories of addiction be generalized to include excessive gambling? *J. Gambl. Behav.* 5, 179–200. doi: 10.1007/BF01024386

Young, K. S. (1998). *Caught in the Net*. New York, NY: JohnWiley.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Psychology

ORIGINAL RESEARCH
published: 12 January 2018
doi: 10.3389/fpsyg.2017.02345

Measurement Invariance of a Classroom Engagement Measure among Academically At-Risk Students

Ryan Glaman [1]* ***and Qi Chen*** [2]

[1] Department of Educational Leadership and Policy Studies, Tarleton State University, Stephenville, TX, United States,
[2] Department of Educational Psychology, University of North Texas, Denton, TX, United States

The current study investigated the measurement invariance of a classroom engagement measure across time points, genders, and ethnicities using a sample of 523 academically at-risk students across grades 7 through 9; this measure was based on Skinner et al.'s (1990) original engagement measure. The engagement measure was comprised of 16 items, yielding three factors: Behavioral Engagement, Behavioral Disaffection, and Emotional Engagement. Configural, metric, and scalar invariance held across the three time points, as did invariance of factor covariances and means, indicating that scores have a similar meaning across all 3 years. The engagement measure also featured adequate configural, metric, and scalar invariance, and invariance of factor covariances and means across genders and ethnicities. These findings suggest the measure is appropriate for investigating substantive hypotheses regarding classroom engagement across different grade levels, genders, and ethnicities. In summary, the current results indicate this measure of classroom engagement is suitable for testing hypotheses regarding group differences in engagement across grade levels, genders, and ethnicities. Researchers may also use this measure to examine relationships between the engagement factors and other important academic outcomes. Limitations of the current study, such as certain caveats regarding convergent validity and internal consistency, are also discussed.

Keywords: engagement, longitudinal data analysis, measurement invariance, multigroup comparison, confirmatory factor analysis

OPEN ACCESS

Edited by:
Jiun-Yu Wu,
National Chiao Tung University, Taiwan

Reviewed by:
Dejan Stevanovic,
Clinic for Neurology and Psychiatry for Children and Youth, Serbia
Joyce Lok Yin Kwan,
The Education University of Hong Kong, Hong Kong

****Correspondence:***
Ryan Glaman
glaman@tarleton.edu

Specialty section:
This article was submitted to Quantitative Psychology and Measurement,
a section of the journal Frontiers in Psychology

Received: *07 March 2017*
Accepted: *22 December 2017*
Published: *12 January 2018*

Citation:
Glaman R and Chen Q (2018) Measurement Invariance of a Classroom Engagement Measure among Academically At-Risk Students. Front. Psychol. 8:2345.
doi: 10.3389/fpsyg.2017.02345

INTRODUCTION

Students' engagement in the classroom is strongly related to academic performance outcomes such as reading performance (Lee, 2014; Lutz Klauda and Guthrie, 2015), mathematics performance (Rimm-Kaufman et al., 2015), and general academic performance (Skinner et al., 1990; Chen et al., 2010). Engagement is also related to other academic variables such as student-teacher relationship quality (Wu et al., 2010) and reading motivation (Lutz Klauda and Guthrie, 2015). Furthermore, failing to engage in the classroom is related to various negative outcomes such as delinquency, substance abuse, and dropout rates (Wang and Fredricks, 2014). Because engagement has relationships with several important academic variables, it is important to consider in both educational research and practice.

One issue with academic engagement as a construct is that there are differences across research studies in terms of its measurement and theoretical definition. For example, some researchers conceptualize engagement as a three-factor construct consisting of behavioral, emotional, and cognitive components (e.g., Burch et al., 2015; Sinatra et al., 2015). Other researchers suggest engagement includes not only behavioral and emotional engagement, but an engagement vs. disaffection component as well; therefore, according to some researchers, engagement is conceptualized as a four-factor construct that includes behavioral engagement, behavioral disaffection, emotional engagement, and emotional disaffection (e.g., Skinner et al., 2008, 2009). Furthermore, engagement may be measured in general, as described above, or it may be domain-specific, such as when measuring reading (Lutz Klauda and Guthrie, 2015) or mathematics engagement (Rimm-Kaufman et al., 2015).

Classroom engagement's theoretical diversity is accompanied by diversity in its measurement as well. Some engagement measures are designed to encapsulate its behavioral, emotional, and cognitive components (Wang and Fredricks, 2014), whereas others attempt to capture emotional and behavioral engagement and disaffection (Skinner et al., 2008). Furthermore, there can also be diversity within a given theoretical perspective; for example, various studies examining emotional and social engagement and disaffection tend to use similar, but slightly different versions of an engagement measure (e.g., Skinner et al., 1998, 2008; Wu et al., 2010).

Despite the complexities with theoretically defining and measuring engagement, the goal of the current study was to use measurement invariance (MI) testing procedures to examine the psychometric properties of a measure of classroom engagement. The measure of interest has been used in empirical research (Chen et al., 2010) and is based on Skinner et al. (1998) measure. This particular measure was chosen because it taps into dimensions of behavioral and emotional engagement and disaffection, theoretical constructs that are well-established in the literature (Skinner et al., 1990, 1998, 2008, 2009) and that predict important outcomes such as academic performance.

Measurement Invariance

Generally speaking, MI testing procedures examine the equivalence of a test's measurement across distinct groups of individuals such as genders or ethnicities. Measurement invariance can be tested using a series of multigroup confirmatory factor analyses (CFAs) that impose increasingly stringent criteria on the model (Cheung and Rensvold, 2002; Millsap, 2011). The first criterion is configural invariance, in which the groups have the same pattern of factor coefficients and "zero-loadings" on the factors; that is, the groups conceptualize the concepts the same way (Vandenberg and Lance, 2000; Cheung and Rensvold, 2002). The next criterion is metric invariance, which examines the equality of the factor loadings across groups. Metric invariance is an important prerequisite for meaningful cross-group comparisons. The third criterion, scalar invariance, refers to the equality of item intercepts across groups. Scalar invariance indicates the latent constructs are measured on the same scale across groups and is necessary for comparing groups' factor means. MI can be assessed not only at the item-level, but at the construct-level as well; for example, the invariance of latent factor means or covariances may be examined across groups. These two tests of MI are typically based on theory and may be used to address substantive research questions (Cheung and Rensvold, 2002).

In general, a test must possess MI across groups in order to make cross-group comparisons on the constructs being measured. While MI can be assessed cross-sectionally, it can also be examined using data gathered longitudinally over multiple occasions (Vandenberg and Lance, 2000). Procedurally, longitudinal MI can be examined using either a multisample approach (i.e., similar to examining cross-sectional MI) or by using an augmented covariance matrix as input (Vandenberg and Lance, 2000). For the current study, the former approach was chosen to avoid the shortcomings associated with the augmented covariance matrix approach, such as increased likelihood of non-convergence and generally worse model fit.

Existing psychometric literature has examined MI for various types of engagement measures, but none have explored the measure derived from Skinner et al. (1998) conceptualization. Some studies have examined MI of measures that include elements of cognitive, affective, and behavioral engagement (e.g., Glanville and Wildhagen, 2007; Wang et al., 2011), observing that these measures are largely invariant across ethnicities and genders. Other studies have tested MI for engagement measures featuring more complex factor structures. For example, Bradshaw et al. (2014) examined the MI of Maryland's Safe and Supportive Schools Initiative survey, which features an engagement measure including six factors: teacher connectedness, student connectedness, academic engagement, whole-school connectedness, culture of equity and fairness, and parent engagement; the authors found this measure was invariant across genders, ethnicities, and grade levels. Other studies have also tested the MI of the Motivation and Engagement Scale, which features five engagement factors: persistence, planning, task management, disengagement, and self-handicapping. Marsh et al. (2011) found this measure was invariant across genders and time points, whereas Martin et al. (2015) showed it was invariant across samples from different countries.

Purpose of the Current Study

Existing psychometric literature on measures of student engagement has yet to examine a measure featuring the theoretical conceptualization described by Skinner et al. (1998), which includes: behavioral engagement, behavioral disaffection, emotional engagement, and emotional disaffection. Therefore, the goal of the current study was to investigate the MI of such an engagement measure longitudinally across students in grades 7 through 9 as well as across ethnicity and gender.

METHOD

Participants

Participants included 523 students attending one of three school districts in Texas (one urban and two small cities). These participants were selected because they were part of a larger longitudinal study investigating the impact of grade retention on academic achievement among at-risk students, in which classroom engagement was also a variable of interest. Participants were recruited across two sequential cohorts in first grade during the fall of 2001 and 2002. Children were eligible to participate in the longitudinal study if they scored below the median score on a state-approved, district-administered measure of literacy, spoke either English or Spanish, were not receiving special education services, and had not previously been retained in first grade. School records identified 1,374 students as being eligible to participate. Because teachers distributed consent forms to parents via children's weekly folders, the exact number of parents who received the consent forms could not be determined. Small gifts to children and the opportunity to win a larger prize in a random drawing were instrumental in obtaining 1,200 returned consent forms, of which 784 parents (65%) provided consent. Analyses on a broad array of archival variables including performance on the district-administered test of literacy, age, gender, ethnicity, eligibility for free or reduced-price lunch, bilingual class placement, cohort, and school context variables (i.e., ethnic composition and percentage of economically disadvantaged students), did not indicate any differences between children with and without consent.

Of these 784 participants, 523 (66.7%) met the inclusion criteria for participation in the current study: they had engagement data from at least one assessment wave, and they were still registered as active in the study at year 9. The sample was 45.0% female and 55.0% male, and its ethnic composition was 37.2% Hispanic, 33.5% White, 25.5% Black, and 3.8% other. A cross-tabulation of ethnic and gender groups is shown in **Table 1**. The results of a chi-square test showed that each of these groups were represented equally within the sample, $\chi^2_{(5)} = 6.282$, $p = 0.280$, Cramer's $V = 0.110$.

Based on attrition analyses, the 523 students in the current sample did not differ from the 261 students who did not complete the study in terms of most demographic variables including: ethnicity, age, socioeconomic status, reading achievement scores based on the Woodcock-Johnson III Broad Reading test (Woodcock et al., 2001), and base-year engagement scores. However, a larger proportion of males remained active in the study than females [$\chi^2_{(1)} = 3.988$, $p = 0.046$, Cramer's $V = 0.071$], a smaller proportion of bilingual students remained active in the study [$\chi^2_{(1)} = 4.615$, $p = 0.032$, Cramer's $V = 0.077$], active students scored slightly higher on the Woodcock-Johnson III Broad Math test than inactive students [$F_{(1,\,754)} = 6.724$, $p = 0.010$, $\eta^2 = 0.009$], and a larger proportion of students whose parents obtained a high school diploma remained active in the study whereas a larger proportion of students whose parents obtained a graduate-level degree dropped out of the study [$\chi^2_{(4)} = 11.173$, $p = 0.025$, Cramer's $V = 0.119$]. However, because the effect sizes for these differences between active and inactive students were small, it was assumed that there were no practical differences between students who dropped out of the study and those who did not.

TABLE 1 | Cross-tabulation of participant demographic frequencies by ethnicity and gender.

Gender	Ethnicity				Row Total
	Hispanic	Black	White	Other	
Female	91	63	70	11	235
Male	104	71	104	9	288
Column total	195	134	174	20	523

This study was carried out in accordance with the recommendations of the Institutional Review Board of Texas A&M University with written informed consent from all participants. All participants gave written informed consent in accordance with the Declaration of Helsinki. The protocol was approved by the Institutional Review Board of Texas A&M University.

Design Overview

Assessments were conducted annually for 9 years, beginning when participants were in the first grade (year 1). Student-report classroom engagement was assessed at years 4, 7, 8, and 9 (these correspond to grades 4, 7, 8, and 9, respectively). However, substantive research suggests developmental differences in engagement exist; that is, students' classroom engagement is likely to shift dramatically between elementary and middle school (Skinner et al., 2008, 2009; Wang et al., 2014). These changes in classroom engagement have been attributed to younger children being developmentally different than young adults in terms their learning strategies, self-regulation, and other factors relevant to classroom engagement (Fredricks and McColskey, 2012; Sinatra et al., 2015). Because fourth grade students are substantially developmentally different from seventh, eighth, and ninth grade students in terms of the cognitive attributes associated with engagement, it would be inappropriate to directly compare those two age groups. Therefore, the year 4 assessment wave was dropped, and only data from years 7, 8, and 9 were included in the MI analyses.

Engagement Measure

Student engagement was measured using a student-report, 18-item scale based on Skinner et al. (1998) original measure. Both English and Spanish versions of the measure were available, and students completed the measure using the language they were more proficient in; ~3.63% of students completed the engagement measure in Spanish at year 7, ~2.68% of students completed the measure in Spanish at year 8, and ~1.91% of students completed the measure in Spanish at year 9. Students indicated how true each item was in describing them using a 1–4 scale (1 = "not at all true," 4 = "very true"). Previous empirical research using this measure of student classroom engagement suggests that it contains three latent factors, Behavioral Engagement, Behavioral Disaffection, and Emotional Engagement, and that one item should be removed from the measure due to low factor loadings (Chen et al., 2010). Example Behavioral Engagement scale items include "When I am in class, I work as hard as I can," and "I try to learn as much as I can about my school subjects." Example Behavioral Disaffection scale items include "When I am in class, I just act like I am working," (reverse scored) and "When I am in class, I just try to look busy" (reverse scored). Example Emotional Engagement scale items include "When I am in class, I feel angry" (reverse scored), and "When I am in class, I feel happy." The internal consistency reliabilities for the current sample across all three time points are shown in **Table 2**. Because dropping an item assessing feeling anxious in class increased the internal consistency of the Emotional Engagement composite scale, one item was dropped, reducing the total number of items in the measure to 16. Overall, the Behavioral Engagement and Behavioral Disaffection scales featured adequate internal consistency reliability across all three measurement occasions according to generally accepted standards (Henson, 2001), but the Emotional Engagement scale did not.

Data Analysis Overview

A series of MI analyses were conducted to examine the longitudinal stability of the engagement measure's structure across 3 years; the configural, metric, and scalar invariance assumptions, as well as invariance of factor covariances and means, were sequentially tested using a CFA framework. Before conducting the MI analyses, the normality of the engagement item responses was examined across the 3 years. All the skewness and kurtosis statistics were within the acceptable range (skewness within ±3.00 and kurtosis within ±8.00; Kline, 2010), indicating all the study variables were normally distributed. We tested the three-factor CFA model from previous research as described above (Chen et al., 2010) with data from the three assessment waves. Model chi-square test statistics, along with Hu and Bentler's (1999) commonly used model fit criteria, were used to provide evidence of adequate model fit. Modification indices were used to provide statistical evidence for the unknown underlying relationships among items. An examination of modification indices indicated that that three pairs of correlated items had consistently large modification indices across all 3 years (i.e., modification index ≥20). Items 2 and 6 both involved concentrating on doing class work. Items 9 and 10 both involved trying to look busy during class. Lastly, items 15 and 17 both involved thinking about non-class-related things during class time. As researchers recommend that use of modification indices have substantive justification (Byrne, 1998; Kline, 2010), the original three-factor CFA was modified to include these three pairs of correlated items due to the pairs' content similarity. This revised three-factor CFA demonstrated an adequate model goodness-of-fit across all three assessment waves, average comparative fit index ($\overline{CFI}$) = 0.952, average Tucker-Lewis index ($\overline{TLI}$) = 0.941, average root-mean-square error of approximation ($\overline{RMSEA}$) = 0.049, average standardized root mean square residual ($\overline{SRMR}$) = 0.050. Fit indices for the three individual assessment waves are shown in **Table 3**.

Additionally, the engagement factors' convergent and discriminant validity were assessed across all three time points by examining the standardized factor pattern/structure coefficients, average variance extracted (AVE) for each factor, and factor correlations. Standardized factor pattern/structure coefficients and associated standard errors across all 3 years are shown in **Table 4** and factor AVEs, correlations, and squared correlations are all shown in **Table 5**. Kline (2010) suggested that standardized factor pattern coefficients of at least 0.70 indicate good convergent validity. While the majority of items across the engagement subscales met or came close to meeting this threshold (see **Table 4**), certain items were problematic across time points, specifically items 13, 14, 15, 16, and 17. Regarding discriminant validity, although the squared factor correlations are relatively high compared to the AVEs in several cases (see **Table 5**), moderately high correlations between these engagement factors have also been observed in prior research (e.g., Skinner et al., 2009), suggesting that the correlations in the present study are consistent with existing theory. Furthermore, Kline (2010) suggested that as long as factor correlations are not excessively high (i.e., ≥0.90 in absolute value), that is evidence of adequate discriminant validity. Therefore, despite relatively low pattern coefficients for select items and high factor correlations in some cases, the current three-factor model demonstrated moderate convergent and discriminant validity across all three time points.

The CFA structure described above was used as the factor structure in testing the longitudinal MI of the engagement measure across the 3 years. To do so, we followed a procedure recommended by Millsap (2011) and employed by previous MI researchers (e.g., Wu and Hughes, 2015). First, we examined the configural invariance of the measurement structures across the three time points. Next, we tested the metric invariance of items' factor loadings by comparing the model for a given year with the model of the previous year(s). This procedure has two advantages: (a) we can know if the parameters remain the same across the 3 years, and (b) we can detect at which years items become non-invariant. For example, first, we tested the metric invariance between year 7 and year 8 ($\boldsymbol{\Lambda}_7 = \boldsymbol{\Lambda}_8$), allowing the factor loadings for year 9 to be freely estimated. Then, we tested the metric invariance between all 3 years ($\boldsymbol{\Lambda}_7 = \boldsymbol{\Lambda}_8 = \boldsymbol{\Lambda}_9$). After confirming the metric invariance assumption, we tested the scalar invariance assumption, the invariance of factor covariances, and the invariance of factor means using the same procedure.

TABLE 2 | Cronbach's alpha reliabilities of the engagement scales across 3 years.

	Behavioral Engagement (7 Items)	Behavioral Disaffection (6 Items)	Emotional Engagement (3 Items)
Year 7	0.810	0.810	0.552
Year 8	0.819	0.795	0.517
Year 9	0.825	0.810	0.572

TABLE 3 | CFA model fit indices across all three measurement occasions.

Year	CFI	TLI	RMSEA	SRMR
7	0.955	0.945	0.048	0.051
8	0.956	0.946	0.047	0.046
9	0.945	0.933	0.053	0.053

TABLE 4 | Standardized factor pattern/structure coefficients (and standard errors) for the CFA model across 3 years.

Item	Year 7	Year 8	Year 9
BehEng			
Item 1	0.668 (0.047)	0.669 (0.046)	0.663 (0.039)
Item 2	0.705 (0.031)	0.731 (0.030)	0.698 (0.030)
Item 5	0.783 (0.028)	0.814 (0.025)	0.823 (0.024)
Item 6	0.738 (0.029)	0.756 (0.028)	0.759 (0.028)
Item 7	0.725 (0.033)	0.721 (0.035)	0.723 (0.030)
Item 13	0.509 (0.048)	0.503 (0.044)	0.413 (0.047)
Item 14	0.430 (0.041)	0.419 (0.041)	0.479 (0.040)
BehDis			
Item 9*	0.656 (0.041)	0.729 (0.034)	0.697 (0.036)
Item 10*	0.692 (0.038)	0.702 (0.041)	0.749 (0.029)
Item 11*	0.668 (0.034)	0.663 (0.037)	0.735 (0.033)
Item 15*	0.566 (0.037)	0.554 (0.033)	0.604 (0.033)
Item 17*	0.497 (0.037)	0.468 (0.041)	0.498 (0.044)
Item 18*	0.786 (0.029)	0.703 (0.047)	0.632 (0.042)
EmoEng			
Item 8*	0.658 (0.056)	0.557 (0.065)	0.722 (0.048)
Item 12*	0.682 (0.056)	0.601 (0.063)	0.632 (0.052)
Item 16	0.438 (0.066)	0.534 (0.056)	0.446 (0.059)

**Denotes reverse-scored items. BehEng, Behavioral Engagement factor; BehDis, Behavioral Disaffection factor; EmoEng, Emotional Engagement factor. Items 3 and 4 are absent from this table because those two items were removed from the measure. Standard errors are shown in parentheses. All pattern coefficients were statistically significant at the p < 0.001 level.*

TABLE 5 | Discriminant validity analyses for the engagement measure CFA model across years 7, 8, and 9.

	Average variance explained	Factor correlations (correlations squared)	
		BehEng	**BehDis**
YEAR 7			
BehEng	0.439	–	
BehDis	0.423	0.819 (0.671)	–
EmoEng	0.363	0.548 (0.300)	0.746 (0.557)
YEAR 8			
BehEng	0.458	–	
BehDis	0.414	0.766 (0.587)	–
EmoEng	0.319	0.711 (0.506)	0.700 (0.490)
YEAR 9			
BehEng	0.443	–	
BehDis	0.443	0.794 (0.630)	–
EmoEng	0.373	0.475 (0.226)	0.685 (0.469)

BehEng, Behavioral Engagement factor; BehDis, Behavioral Disaffection factor; EmoEng, Emotional Engagement factor. All correlations were statistically significant at the p < 0.001 level.

To evaluate the longitudinal MI and compare the fit of the individual models, we used the chi-square difference test ($\Delta\chi^2$; Kline, 2010). However, because the chi-square test is highly sensitive to sample size, we also examined other overall and incremental indicators of model fit. We examined several overall indicators of model fit, such as the CFI, Tucker Lewis Index (TLI), root mean square error of approximation (RMSEA) and SRMR. Hu and Bentler (1999) suggested that values >0.95 for the CFI and TLI, values <0.06 for the RMSEA, and values <0.08 for the SRMR indicate good overall model fit. Though these are not intended to be hard-and-fast cutoff values, they can be used to guide interpretation of model fit. We also used two model fit indicators to examine the incremental changes in model fit across the longitudinal MI analyses, including change in the comparative fit index (ΔCFI) and the Tucker-Lewis index (ΔTLI); when $\Delta\text{CFI} \leq 0.02$ (Cheung and Rensvold, 2002) and $\Delta\text{TLI} \leq 0.05$ (Little, 1997), then the two comparative models are not substantially different from one another. Measurement invariance researchers suggest examining a variety of overall and incremental indicators when interpreting model fit and change (Vandenberg and Lance, 2000).

Following the longitudinal MI analyses, we conducted multigroup comparisons to examine whether the MI assumptions held across genders and ethnic groups. We investigated three different ethnic groups: Black, Hispanic, and White, and all three measurement occasions were accounted for in these MI analyses.

RESULTS

The current data featured a nested structure (students nested within classrooms). This nesting structure was accounted for using the TYPE = COMPLEX routine in Mplus version 6.11 with the robust standard error estimator (Muthén and Muthén, 2010). Overall, 6.39% of the data were missing and had properties in line with the missing at random (MAR) condition according to missing data analyses. Therefore, participants with scores on at least one assessment wave were included in the analysis, and missing data were handled using multiple imputation. Ten datasets were imputed and the results described in the current paper represent the overall results pooled across all 10 imputations.

Longitudinal MI

Results indicated the revised model of the engagement measure featuring three latent constructs and three sets of correlated items adequately fit the longitudinal data, $\chi^2_{(294)} = 705.406$, $p < 0.001$, CFI = 0.942, TLI = 0.929, RMSEA = 0.052, SRMR = 0.052. These values are also shown in **Table 6** as Model 1.1. This indicates the configural invariance assumption held for the 3-year data; all three time points had the same pattern of factor coefficients.

Results of the metric longitudinal MI tests are shown in **Table 6** as Model 2.1 and Model 2.2. Model fit indices indicate both models fit the data adequately. Furthermore, all model change statistics were smaller than the suggested critical values, indicating that differences in fit between the two models were statistically negligible. Since Model 2.2, the more restricted model that assumed the pattern of factor loadings was identical across time points, fit the data just as well as previous models, metric invariance was established for the engagement measure.

Results of the scalar longitudinal MI analyses are also shown in **Table 6** as Model 3.1 and Model 3.2. As with previous models, fit

TABLE 6 | Model fit test statistics, fit indices, and their changes for the longitudinal MI analyses of years 7, 8, and 9.

	Model fit test statistics and fit indices						**Change of model fit test statistics and fit indices**			
	χ^2	***df***	**CFI**	**TLI**	**RMSEA**	**SRMR**	**$\Delta\chi^2$**	**Δ*df***	**ΔCFI**	**ΔTLI**
CONFIGURAL INVARIANCE										
1.1	705.406	294	0.942	0.929	0.052	0.052	–	–	–	–
METRIC INVARIANCE										
2.1 $\boldsymbol{\Lambda}_7 = \boldsymbol{\Lambda}_8$	722.772	307	0.941	0.931	0.051	0.053	**17.366**	13	**0.001**	**−0.002**
2.2 $\boldsymbol{\Lambda}_7 = \boldsymbol{\Lambda}_8 = \boldsymbol{\Lambda}_9$	739.208	320	0.941	0.934	0.050	0.056	**16.436**	13	**<0.001**	**−0.003**
SCALAR INVARIANCE										
3.1 $\boldsymbol{\tau}_7 = \boldsymbol{\tau}_8$	773.840	333	0.938	0.933	0.050	0.058	34.632	13	**0.003**	**0.001**
3.2 $\boldsymbol{\tau}_7 = \boldsymbol{\tau}_8 = \boldsymbol{\tau}_9$	817.059	346	0.934	0.931	0.051	0.060	43.219	13	**0.004**	**0.002**
FACTOR COVARIANCES INVARIANT										
4.1 $\boldsymbol{\Phi}_7 = \boldsymbol{\Phi}_8$	827.766	349	0.932	0.930	0.051	0.062	10.707	3	**0.002**	**0.001**
4.2 $\boldsymbol{\Phi}_7 = \boldsymbol{\Phi}_8 = \boldsymbol{\Phi}_9$	833.481	352	0.932	0.931	0.051	0.064	**5.715**	3	**<0.001**	**−0.001**
FACTOR MEANS INVARIANT										
5.1 $\boldsymbol{\mu}_7 = \boldsymbol{\mu}_8$	844.293	355	0.931	0.930	0.051	0.064	10.812	3	**0.001**	**0.001**
5.2 $\boldsymbol{\mu}_7 = \boldsymbol{\mu}_8 = \boldsymbol{\mu}_9$	862.853	358	0.929	0.928	0.052	0.066	18.56	3	**0.002**	**0.002**

Bold font indicates the difference between two comparative models is statistically negligible. Λ, factor loading matrix; τ, item intercept vector; Φ, factor covariance matrix; μ, factor mean vector; subscripts indicate the years measurements were collected; df, degrees of freedom; CFI, comparative fit index; TLI, Tucker-Lewis index; RMSEA, root mean square error of approximation; SRMR, standardized root mean square residual.

statistics indicate both models fit the data adequately. While the $\Delta\chi^2$ model fit test statistics were statistically significant for both models, the ΔCFI and ΔTLI both indicate the models fit the data just as well as the previous models. Therefore, the assumption of scalar invariance held.

Results of the invariance analyses for the factor covariances are shown in **Table 6** as Model 4.1 and 4.2. Similar to previous models, fit statistics show that both models fit the data adequately; both the ΔCFI and ΔTLI indicate that these models fit the data as well as the previous models, suggesting that covariances among the three factors are equivalent across time points. Factor correlations across all three time points for Model 4.2 are shown in **Table 7**. Please note that, although the covariances between the latent constructs were equivalent across time, these correlations differ slightly across all three time points because the latent factors featured slightly different variances.

Lastly, results of the invariance analyses for the factor means are also shown in **Table 6** as Model 5.1 and Model 5.2. As with previous models, fit statistics indicate that both models adequately fit the data, and the ΔCFI and ΔTLI both suggest these models fit the data as well as previous models. Therefore, the means of the latent factors were assumed to be equal across time points. In sum, the longitudinal MI analyses indicated the engagement measure featured configural, metric, and scalar invariance, as well as equivalence of latent factor covariances and means across all three time points.

Measurement Invariance across Gender and Ethnicity

We also used MI testing procedures to examine the MI of the engagement measure across gender and ethnic groups; results of these analyses are shown in **Table 8**. Note that all three measurement occasions were accounted for in these analyses. For gender, models examining configural, metric, and scalar invariance, and invariance of factor covariances and means, all fit the data adequately based on model fit indices. Furthermore, although most $\Delta\chi^2$ tests indicated certain models were statistically significantly different from one another in terms of overall model fit, the ΔCFI and ΔTLI both indicated the more constrained models all fit the data just as well as the previous models. Two $\Delta\chi^2$ tests produced negative values, indicating that these difference tests cannot be interpreted and used to test for statistically significant differences in model fit. Therefore, we used the Wald test of parameter constraints to examine differences in model fit for these two comparisons. The model featuring metric invariance did not statistically significantly differ from the model featuring configural invariance in terms of overall fit, Wald $\chi^2_{(13)} = 8.484$, $p = 0.811$, nor did the model featuring invariant factor covariances differ from the model featuring scalar invariance, Wald $\chi^2_{(3)} = 3.102$, $p = 0.376$. In sum, based on the overall fit statistics and the ΔCFI and ΔTLI, we concluded the engagement measure featured adequate configural, metric, and scalar invariance, as well as equivalence of factor covariances and means across males and females.

Regarding ethnicity, model fit indices suggest the five models testing the configural, metric, and scalar invariance, and invariance of latent factor covariances and means, of the engagement measure all fit the data reasonably well (see **Table 8**). Although some $\Delta\chi^2$ tests suggested that some models fit the data statistically significantly differently from one another, the ΔCFI and ΔTLI both indicated that the more constrained models fit the data as well as the previous models that feature fewer model constraints. Thus, the configural, metric, and scalar invariance assumptions held across ethnic groups, as did invariance of factor covariances and means.

DISCUSSION

Engagement Measure MI

The current results indicated that the classroom engagement measure featured adequate configural, metric, and scalar invariance across time points, genders, and ethnicities. These results suggest that scores on the engagement measure have approximately the same meaning across these groups, and that this measure is appropriate for use when testing substantive hypotheses regarding developmental changes between grades 7 and 9, as well as gender and ethnic differences in engagement among students within this grade range.

The present results tie in well with previous literature examining the MI of other classroom engagement measures. Existing research has shown that other measures based on different theoretical conceptualizations of engagement are also invariant across groups. For example, engagement measures featuring cognitive, affective, and behavioral components were found to be invariant across genders and ethnic groups (Glanville and Wildhagen, 2007; Wang et al., 2011). Research on more complex engagement measures, such as Maryland's Safe and Supportive Schools Initiative Survey (Bradshaw et al., 2014) and the Motivation and Engagement Scale (Marsh et al., 2011), have also demonstrated these measures' invariance across various groups such as genders and grade levels. The current study's findings add to the psychometric literature on classroom engagement measures, demonstrating that this measure, which is based on Skinner et al. (2008) theoretical conceptualization of engagement, is also invariant across grade levels, ethnic groups, and genders. Therefore, the engagement measure examined in the current study is an additional measure, stemming from a different theoretical perspective on engagement that may be used in substantive research to explore cross-group comparisons in behavioral engagement, behavioral disaffection, and emotional engagement.

Furthermore, the current findings also indicated that the factor means and covariances were invariant across time points, ethnic groups, and genders, suggesting that average levels of behavioral engagement, behavioral disaffection, emotional engagement, and the relationships between them remained consistent across these groups. The current findings align with those from previous research regarding factor covariances; existing research has shown that relationships among behavioral engagement and disaffection, and emotional engagement were invariant across grade levels (Skinner et al., 2008, 2009) and genders (Skinner et al., 2009). However, the results regarding invariant factor means run counter to those observed in prior engagement research.

Regarding grade level-related changes in engagement, research on elementary and middle school students has shown that elementary students tend to have higher emotional and behavioral engagement than middle school students (Skinner et al., 2008, 2009); additional research suggests that classroom disengagement increases between elementary and middle school (Wang et al., 2014). That said, the lack of change in engagement across grade levels in the current study may be because changes in classroom engagement occur primarily between elementary and middle school. In past research, changes

TABLE 7 | Factor correlations across 3 years for model 4.2.

	BehEng with BehDis	BehEng with EmoEng	BehDis with EmoEng
Year 7	0.798	0.524	0.678
Year 8	0.754	0.580	0.732
Year 9	0.775	0.523	0.711

BehEng, Behavioral Engagement factor; BehDis, Behavioral Disaffection factor; EmoEng, Emotional Engagement factor. All correlations were statistically significant at the p < 0.001 level.

TABLE 8 | Model fit test statistics, fit indices, and their changes for the MI analyses for gender and ethnicity.

	Model fit test statistics and fit indices						Change of model fit test statistics and fit indices			
	χ^2	df	CFI	TLI	RMSEA	SRMR	$\Delta\chi^2$	Δdf	ΔCFI	ΔTLI
GENDER										
Configural	585.500	196	0.942	0.929	0.050	0.050	–	–	–	–
Metric $\boldsymbol{\Lambda}_M = \boldsymbol{\Lambda}_F$	585.024	209	0.944	0.935	0.048	0.052	−0.476*	13	**−0.002**	**−0.006**
Scalar $\boldsymbol{\tau}_M = \boldsymbol{\tau}_F$	630.885	222	0.939	0.934	0.048	0.053	45.861	13	**0.005**	**0.001**
Factor Covariances $\Phi_M = \Phi_F$	630.323	225	0.939	0.935	0.048	0.058	−0.562*	3	**<0.001**	**−0.001**
Factor Means $\boldsymbol{\mu}_M = \boldsymbol{\mu}_F$	650.970	228	0.937	0.933	0.049	0.070	20.647	3	**0.002**	**0.002**
ETHNICITY										
Configural	648.856	294	0.946	0.934	0.049	0.051	–	–	–	–
Metric $\boldsymbol{\Lambda}_B = \boldsymbol{\Lambda}_H = \boldsymbol{\Lambda}_W$	682.103	320	0.945	0.938	0.047	0.060	33.247	26	**0.001**	**−0.004**
Scalar $\boldsymbol{\tau}_B = \boldsymbol{\tau}_H = \boldsymbol{\tau}_W$	781.222	346	0.934	0.931	0.050	0.066	99.199	26	**0.011**	**0.007**
Factor Covariances $\boldsymbol{\Phi}_B = \boldsymbol{\Phi}_H = \boldsymbol{\Phi}_W$	789.592	352	0.934	0.932	0.050	0.073	**8.37**	6	**<0.001**	**−0.001**
Factor Means $\boldsymbol{\mu}_B = \boldsymbol{\mu}_H = \boldsymbol{\mu}_W$	832.159	358	0.928	0.928	0.051	0.084	42.567	6	**0.006**	**0.004**

*Bold font indicates the difference between two comparative models is statistically negligible. Λ, factor loading matrix; τ, item intercept vector; Φ, factor covariance matrix; μ, factor mean vector; subscripts indicate the groups of the measures collected; df, degrees of freedom; CFI, comparative fit index; TLI, Tucker-Lewis index; RMSEA, root mean square error of approximation; SRMR, standardized root mean square residual; M, male; F, female; B, Black; H, Hispanic; W, White. *Denotes the negative test statistic cannot be interpreted, and a Wald test of parameter constraints was examined instead.*

in classroom engagement have been attributed to younger children being developmentally different than older children in terms their learning strategies, self-regulation, and other cognitive factors related to classroom engagement (Fredricks and McColskey, 2012; Sinatra et al., 2015). Because the current student sample is older than those examined in prior research, engagement levels may have stabilized by the time students reached grade 7 and remained consistent throughout grades 7, 8, and 9.

Regarding gender, previous research indicates that girls tend to be higher in behavioral and emotional engagement than boys (Skinner et al., 2008, 2009; Wang et al., 2011, 2014); Wang and Fredricks (2014) also observed that girls were higher in cognitive engagement than boys. It is unknown at this time why the current results do not align with those from previous research. However, due to the nature of the current sample being composed of lower-achieving students, there may have been additional variables at play that impacted the present findings that may not otherwise be present in other samples. Previous research has shown that outside variables, such as teacher-student interaction quality (Rimm-Kaufman et al., 2015) do interact with how gender relates to engagement; it is possible that such variables played a role in the current study, but were not accounted for.

Lastly, prior research on ethnicity suggests that White students tend to have higher behavioral engagement and lower emotional engagement than Black students (Wang et al., 2011; Wang and Fredricks, 2014). Once again, it is unknown why the current results do not match those from past studies. As with the issue described above regarding the lack of gender differences, other variables associated with the low-achieving sample makeup may have played a role in the current results.

Limitations and Future Directions

One limitation of the current study is that the sample was selected based on students who scored below the median on a district-administered literacy measure. Therefore, the current results may apply only to lower-achieving students and not to normally- or higher-achieving students. A second limitation was that the engagement measure was administered at four different time points, only three of which were used in the current study. Although the current analyses indicated the measurement of the classroom engagement measure was consistent across the 3-year period, because it was not administered at a larger number of time points, it is impossible to know how well the longitudinal MI would hold over a longer period of time. Future research is needed to examine the measurement properties of this engagement measure in both more diverse samples, and over longer periods of time.

Furthermore, the three-factor CFA model that was examined features some caveats that should be accounted for when using and interpreting this engagement measure. First, the Emotional Engagement subscale featured relatively poor internal consistency reliability compared to the other two subscales (see **Table 2**). Although Emotional Engagement's internal consistency was low in this particular study, reliability estimates can vary between different samples and test administrations (Henson, 2001). Therefore, the low reliability observed in the current study may be due to the nature of the sample that was studied. Future researchers employing this classroom engagement measure should examine the reliability of all three subscales, which would help identify whether the low internal consistency in the current study was an anomaly or part of a broader pattern. Also, although the current CFA model fit the data well overall, it featured only moderate convergent and discriminant validity. Future researchers should bear in mind these slight validity limitations and take note that the engagement subscales are highly related to one another, as shown both in the current study, and in previous research (Skinner et al., 2009).

CONCLUSION

The three-factor engagement measure examined in the current study, derived from Skinner et al.'s (2008) theoretical conceptualization, features adequate configural, metric, and scalar MI, as well as equivalence of factor covariances and means, across grade levels, genders, and ethnicities; our results support the psychometric consistency of the engagement measure across these three variables.

Therefore, based on the current study's findings, this measure of classroom engagement is suitable for testing hypotheses regarding group differences in engagement across genders and ethnicities, as well as for studying grade level-related changes in engagement. Given the stability of this measure across genders, ethnicities, and grades, researchers may also use it to examine relationships between the engagement factors and other important academic outcomes.

AUTHOR CONTRIBUTIONS

RG: conducted the data analyses and interpretation, performed majority of the writing, and edited the final paper. QC: developed the original research idea, gained access to the data, and guided the overall research process.

ACKNOWLEDGMENTS

We would like to thank Dr. Jan Hughes for providing us with access to the Project Achieve data and guidance on data questions.

REFERENCES

Bradshaw, C. P., Waasdorp, T. E., Debnam, K. J., and Johnson, S. L. (2014). Measuring school climate in high schools: a focus on safety, engagement, and the environment. *J. Sch. Health* 84, 593–604. doi: 10.1111/josh.12186

Burch, G. F., Heller, N. A., Burch, J. J., Freed, R., and Steed, S. A. (2015). Student engagement: developing a conceptual framework and survey instrument. *J. Educ. Bus.* 90, 224–229. doi: 10.1080/08832323.2015.1019821

Byrne, B. M. (1998). *Structural Equation Modeling with LISREL, PRELIS, and SIMPLIS: Basic Concepts, Applications, and Programming.* Mahwah, NJ: Lawrence Erlbaum.

Chen, Q., Hughes, J. N., Liew, J., and Kwok, O. (2010). Joint contributions of peer acceptance and peer academic reputation to achievement in academically at-risk children: mediating processes. *J. Appl. Dev. Psychol.* 31, 448–459. doi: 10.1016/j.appdev.2010.09.001

Cheung, G. W., and Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Struct. Equ. Model.* 9, 233–255. doi: 10.1207/S15328007SEM0902_5

Fredricks, J. A., and McColskey, W. (2012). "The measurement of student engagement: A comparative analysis of various methods and student self-report instruments," in *Handbook of Research on Student Engagement,* eds S. L. Christenson, A. L. Reschly, and C. Wylie (New York, NY: Springer), 763–782.

Glanville, J. L., and Wildhagen, T. (2007). The measurement of school engagement: assessing dimensionality and measurement invariance across race and ethnicity. *Educ. Psychol. Meas.* 67, 1019–1041. doi: 10.1177/0013164406299126

Henson, R. K. (2001). Understanding internal consistency reliability estimates: a conceptual primer on coefficient alpha. *Meas. Eval. Counsel. Dev*. 34, 177–189.

Hu, L., and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struc. Equ. Model.* 6, 1–55. doi: 10.1080/10705519909540118

Kline, R. B. (2010). *Principles and Practice of Structural Equation Modeling, 3rd Edn*. New York, NY: Guilford.

Lee, J. (2014). The relationship between student engagement and academic performance: is it a myth of reality? *J. Educ. Res.* 107, 177–185. doi: 10.1080/00220671.2013.807491

Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: practical and theoretical issues. *Multivariate Behav. Res.* 32, 53–76. doi: 10.1207/s15327906mbr3201_3

Lutz Klauda, S., and Guthrie, J. T. (2015). Comparing relations of motivation, engagement, and achievement among struggling and advanced adolescent readers. *Read. Writ.* 28, 239–269. doi: 10.1007/s11145-014-9523-2

Marsh, H. W., Liem, G. A. D., Martin, A. J., Morin, A. J. S., and Nagengast, B. (2011). Methodological measurement fruitfulness of exploratory structural equation modeling (ESESM): new approaches to key substantive issues in motivation and engagement. *J. Psychol. Assess.* 29, 322–346. doi: 10.1177/0734282911406657

Martin, A. J., Yu, K., Papworth, B., Ginns, P., and Collie, R. J. (2015). Motivation and engagement in the United States, Canada, United Kingdom, Australia, and China: testing a multi-dimensional framework. *J. Psychol. Assess.* 32, 103–114. doi: 10.1177/0734282914546287

Millsap, R. E. (2011). *Statistical Approaches to Measurement Invariance.* New York, NY: Routledge.

Muthén, L. K., and Muthén, B. O. (2010). *Mplus User's Guide, 6th Edn.* Los Angeles, CA: Muthén and Muthén.

Rimm-Kaufman, S. E., Baroody, A. E., Larsen, R. A. A., Curby, T. W., and Abry, T. (2015). To what extend do teacher-student interaction quality and student gender contribute to fifth graders' engagement in mathematics learning? *J. Educ. Psychol.* 107, 170–185. doi: 10.1037/a0037252

Sinatra, G. M., Heddy, B. C., and Lombardi, D. (2015). The challenges of defining and measuring student engagement in science. *Educ. Psychol.* 50, 1–13. doi: 10.1080/00461520.2014.1002924

Skinner, E. A., Wellborn, J. G., and Connell, J. P. (1990). What it takes to do well in school and whether I've got it: a process model of perceived control and children's engagement and achievement in school. *J. Educ. Psychol.* 82, 22–32.

Skinner, E. A., Zimmer-Gembeck, M. J., Connell, J. P., (1998). Individual differences and the development of perceived control. *Monogr. Soc. Res. Child Dev*. 63 i–vi, 1–220. doi: 10.2307/1166220

Skinner, E., Furrer, C., Marchand, G., and Kindermann, T. (2008). Engagement and disaffection in the classroom: Part of a larger motivational dynamic? *J. Educ. Psychol.* 100, 765–781. doi: 10.1037/a0012840

Skinner, E., Kindermann, T. A., and Furrer, C. J. (2009). A motivational perspective on engagement and disaffection: conceptualization and assessment of children's behavioral and emotional participation in academic activities in the classroom. *Educ. Psychol. Meas.* 69, 493–525. doi: 10.1177/0013164408323233

Vandenberg, R. J., and Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. *Organ. Res. Methods* 3, 4–70. doi: 10.1177/109442810031002

Wang, Z., Bergin, C., and Bergin, D. A. (2014). Measuring engagement in fourth to twelfth grade classrooms: the classroom engagement inventory. *Sch. Psychol. Q.* 29, 517–535. doi: 10.1037/spq0000050

Wang, M. T., and Fredricks, J. A. (2014). The reciprocal links between school engagement, youth problem behaviors, and school dropout during adolescence. *Child Dev*. 85, 722–737. doi: 10.1111/cdev.12138

Wang, M. T., Willett, J. B., and Eccles, J. S. (2011). The assessment of school engagement: examining dimensionality and measurement invariance by gender and race/ethnicity. *J. Sch. Psychol.* 49, 465–480. doi: 10.1016/j.jsp.2011.04.001

Woodcock, R. W., McGrew, K. S., and Mather, N. (2001). *WJ-III Tests of Achievement*. Itasca, IL: Riverside.

Wu, J.-Y., and Hughes, J. N. (2015). Teacher Network of Relationships Inventory: measurement invariance of academically at-risk students across ages 6 to 15. *School Psychol. Q.* 30, 23–36. doi: 10.1037/spq0000063

Wu, J. Y., Hughes, J. N., and Kwok, O. M. (2010). Teacher-student relationship quality type in elementary grades: effects on trajectories for achievement and engagement. *J. Sch. Psychol.* 48, 357–387. doi: 10.1016/j.jsp.2010.06.004

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

METHODS
published: 24 August 2017
doi: 10.3389/fpsyg.2017.01455

Applying Permutation Tests and Multivariate Modification Indices to Configurally Invariant Models That Need Respecification

Terrence D. Jorgensen *

Research Institute for Child Development and Education, University of Amsterdam, Amsterdam, Netherlands

The assumption of equivalence between measurement-model configurations across groups is typically investigated by evaluating overall fit of the same model simultaneously to multiple samples. However, the null hypothesis (H_0) of configural invariance is distinct from the H_0 of overall model fit. Permutation tests of configural invariance yield nominal Type I error rates even when a model does not fit perfectly (Jorgensen et al., 2017, in press). When the configural model requires modification, lack of evidence against configural invariance implies that researchers should reconsider their model's structure simultaneously across all groups. Application of multivariate modification indices is therefore proposed to help decide which parameter(s) to free simultaneously in all groups, and I present Monte Carlo simulation results comparing their Type I error control to traditional 1-*df* modification indices. I use the Holzinger and Swineford (1939) data set to illustrate these methods.

Keywords: configural invariance, permutation tests, measurement equivalence/invariance, confirmatory factor analysis, Lagrange multipliers, modification indices

OPEN ACCESS

Edited by:
Oi-Man Kwok,
Texas A&M University, United States

Reviewed by:
Heungsun Hwang,
McGill University, Canada
Eun Sook Kim,
University of South Florida,
United States

****Correspondence:***
Terrence D. Jorgensen
t.d.jorgensen@uva.nl

Specialty section:
This article was submitted to
Quantitative Psychology and
Measurement,
a section of the journal
Frontiers in Psychology

Received: *15 May 2017*
Accepted: *11 August 2017*
Published: *24 August 2017*

Citation:
Jorgensen TD (2017) Applying
Permutation Tests and Multivariate
Modification Indices to Configurally
Invariant Models That Need
Respecification.
Front. Psychol. 8:1455.
doi: 10.3389/fpsyg.2017.01455

Many behavioral researchers do not have the luxury of being able to directly observe the phenomena they study. For example, organizational researchers need to measure job satisfaction or morale. Clinicians need to measure various psychological disorders. Social psychologists and sociologists need to measure attitudes and social orientations. Educational researchers need to measure teaching and learning outcomes. Often, researchers rely on indirect measures, such as self-report scales, and psychometric tools, such as reliability estimates and latent trait models [e.g., confirmatory factor analysis (CFA) and item-response theory (IRT) models] facilitate evaluation of the quality of those measurements.

Similarly frequent is the need for researchers to compare groups, in either experimental (e.g., treated vs. control) or observational contexts (e.g., demographic or intact groups). In order to make valid comparisons of scale responses across groups, the scale must function equivalently for those groups. In other words, if measurement parameters are equivalent across groups, observed group means will only differ as a function of differences on the latent trait itself (Meredith, 1993). Measurement equivalence/invariance (ME/I) has received a great deal of attention in the methodological literature, so I provide only a cursory introduction here; interested readers are encouraged to find more in-depth discussion in Meredith (1993); Reise et al. (1993); Vandenberg and Lance (2000), and Putnick and Bornstein (2016).

Latent trait models facilitate the investigation of ME/I, and different levels of ME/I have been defined according to categories of model parameters. In a CFA framework, configural invariance is represented in a model with the same pattern of fixed and free (i.e., near-zero and substantial) factor loadings across groups, although the values of these parameters may differ across groups. When fitting models to multivariate normally distributed data using maximum likelihood estimation, the null hypothesis (H_0) of configural invariance is traditionally tested using a likelihood-ratio test statistic (LRT)[1], which is distributed as a χ^2 random variable with *df* equal to the number of nonredundant observed means and (co)variances minus the number of estimated model parameters. Configural invariance is the least restrictive level of ME/I, so it can be used as a baseline model for comparing more restrictive assumptions of ME/I, which are represented by models that are nested within the configural model.

Metric equivalence (or "weak" invariance) indicates the additional assumption that the values of factor loadings are equal across groups, and this assumption must hold in order to make valid across-group comparisons of latent variances or correlations. This model is nested within the configural model, so a $\Delta\chi^2$ test can be used to test the H_0 of exact metric equivalence. If a researcher concludes that full (or partial[2]) metric equivalence holds, that model is used as a baseline model to test scalar equivalence (or "strong" invariance") by additionally constraining indicator intercepts (or thresholds for binary or ordinal indicators) to equality across groups. Scalar invariance is required for valid comparisons of latent means to be made. Researchers can also test homogeneity of residual variances across groups ("strict" invariance), but because that assumption is not required for valid comparisons of latent parameters, it is not tested as often (Putnick and Bornstein, 2016).

The current paper discusses recent advances only in tests of configural invariance, which is the least restrictive level of invariance. A false H_0 would imply that model configurations differ across groups, in which case data-generating population processes do not share all the same parameters across groups. A test that rejects the H_0 of configural invariance would therefore prohibit researchers from testing more restrictive levels of ME/I. Currently, configural invariance is assessed by evaluating the overall fit of the configural model (Putnick and Bornstein, 2016). A significant LRT or fit indices that do not meet criteria for adequate fit (Hu and Bentler, 1999) would be expected when configural invariance does not hold across populations, because the hypothesized model could only represent the data-generating process for one (subset of) group(s), and would be incorrect for at least one of the other groups; thus, the poor fit of the model to that group would be reflected in the overall model fit measures. However, the fact that a false H_0 should lead to a poor fit does not imply the reverse[3]: If the model fits poorly, that does not necessarily imply that true population models are configurally noninvariant. A hypothesized configural model could fit poorly for a different reason; specifically, the true data-generating process might be equivalent across groups (i.e., H_0 of configural invariance is true), but the specified model is a poor approximation of the true functional form of that process (i.e., false H_0 that the model is correctly specified).

Using a newly proposed permutation test of configural invariance (Jorgensen et al., 2017, in press), the H_0 of configural invariance can be tested with nominal Type I error rates even when the H_0 of correct specification is false. I extend this line of research by proposing the use of multivariate modification indices (Bentler and Chou, 1992) to guide researchers in respecifying their inadequately fitting configural models when there is no evidence against the H_0 of group equivalence in true model configurations. This study is therefore only concerned with the situation when the H_0 of configural invariance is true (but the model does not fit well), not when the H_0 is false. To evaluate the use of multivariate modification indices for the purpose of testing whether the same parameter should be freed simultaneously across groups, I designed a small-scale simulation study as a proof of concept to show that they are capable of preventing Type I error inflation better than traditional 1-*df* modification indices, which test parameters in only one group at a time rather than simultaneously across all groups.

I begin by reviewing in more detail issues with testing model fit vs. configural invariance, using an analysis of the classic Holzinger and Swineford (1939) dataset to demonstrate the use of the permutation test and to illustrate the implication of a configurally invariant model that requires respecification. I then introduce Bentler and Chou's (1992) multivariate extension of modification indices, which are recently available in the open-source `lavaan` package (Rosseel, 2012) for structural equation modeling (SEM) in R (R Core Team, 2017), and discuss how they can be used in the context of respecifying a multigroup model in a way consistent with the H_0 of configural invariance. I then describe the small-scale Monte Carlo simulation study comparing Type I error rates using univariate and multivariate modification indices. I conclude with recommendations for future applied and methodological research.

ISSUES WITH MODEL-FIT TESTS OF CONFIGURAL INVARIANCE

Configural invariance in a multigroup context is equivalence in model configurations across the populations of interest. The analysis models are typically specified as configurally invariant, and the LRT of overall model fit is used to evaluate whether the model adequately approximates the population models. As noted

[1] Although the configural model is only tested with a single model's χ^2 statistic, this statistic is nonetheless equal to -2 times the difference between log-likelihoods of models representing two competing hypotheses: the hypothesized configural model (labeled H0 in the output from software such as M*plus* and lavaan) and the saturated model (labeled H1, representing the default alternative hypothesis of a completely unrestricted model). Because the saturated model has χ^2 and $df = 0$, a $\Delta\chi^2$ test between the configural and saturated models would therefore be calculated by subtracting zero from the configural model's χ^2 and *df*, yielding the same values.

[2] Partial invariance models posit that some, but not all, measurement parameters can be constrained to equality across groups or occasions, which still allows valid comparisons of latent parameters across groups Byrne et al., 1989.

[3] This logical fallacy is referred to as *affirming the consequent*, and has the general form: A implies B, B is true, therefore A is true. This is demonstrably invalid using simple examples for which it is false, such as: "If today is Saturday, it is the weekend. It is in fact the weekend; therefore it is Saturday." The fact that it is the weekend does not imply it is Saturday because it could also be Sunday; there are multiple conditions that could lead to the same state.

in the Introduction, rejection of the H_0 of exact model fit could imply numerous conditions, including but not limited to the following: (a) the hypothesized model corresponds well to one or more populations but poorly to at least one other; (b) the model does not correspond to any group's model, for different reasons across groups; (c) all groups true models are configurally invariant, but the hypothesized model does not correspond to that shared functional form. Thus, when a model's overall fit to multiple groups needs improvement, the decision of how to respecify the model would depend on which condition led to poor overall fit.

Because the LRT is a test of overall exact fit of the model to the data, two potential sources of misspecification are confounded (Cudeck and Henly, 1991; MacCallum, 2003): estimation discrepancy (due to sampling error) and approximation discrepancy (due to a lack of correspondence between the population and analysis models). Because configural invariance is assessed by testing the absolute fit of the configural model, the LRT for a multigroup model further confounds two sources of approximation discrepancy (Jorgensen et al., 2017, in press): the overall discrepancy between population and analysis models could be partitioned into (a) differences between groups' true population models and (b) discrepancies between each group's population and analysis models. The H_0 of configural invariance only concerns the former source of approximation discrepancy (which I will refer to as *group discrepancy*), whereas the latter source is an issue of model-fit in general (which I will refer to as *overall approximation discrepancy*).

Good model fit and equivalent model configurations are both important foundational assumptions of ME/I because testing equality of measurement parameters is only valid if the estimated parameters correspond to actual parameters of the true data-generating process. But merely testing the overall fit of a configural model does not provide adequate information about whether model configurations can be assumed equivalent across groups. It is possible (perhaps even probable) that a model provides as good a description of one population as it does for another population (e.g., men and women or respondents from different countries), even if the model fits poorly or only approximately well. Evaluating overall fit therefore tests the wrong H_0 by confounding group equivalence and overall exact model fit into a single test. The permutation method introduced by Jorgensen et al. (2017, in press) disentangles group discrepancy from overall approximation discrepancy.

Another common issue with model-fit evaluation is the common perception that the LRT nearly always rejects good models because SEM requires large sample sizes for estimation. Although it is true that power is a function of sample size, an analysis model that corresponds perfectly with a true population model would not yield inflated Type I errors (actually, small-sample bias would; Nevitt and Hancock, 2004) because the H_0 would be true. But because theoretical models are more realistically interpreted as approximations to more complex population models (MacCallum, 2003), the H_0 of exact fit should rarely be expected to be precisely true in practice. In order to help researchers evaluate the degree to which a H_0 is false, numerous indices of approximate fit have been proposed since the 1970s, analogous to providing standardized measures of effect size that accompany a null-hypothesis significance test in other contexts (e.g., Cohen's *d* to accompany a *t*-test result).

Unfortunately, approximate fit indices (AFIs) or their differences (Δ) between competing models rarely have known sampling distributions. Even when they do [e.g., the root mean-squared error of approximation (RMSEA); Steiger and Lind, 1980], it is often unclear how to interpret the magnitude of a (Δ)AFI. Researchers frequently rely on rule-of-thumb cutoffs, such as those proposed by Hu and Bentler (1999) for AFIs or by Cheung and Rensvold (2002) for ΔAFIs, either based on intuition or derived from simulation studies under specific conditions that might not generalize to the wide array of SEMs encountered in practice. Although it is reasonable to argue that models with only negligible misspecifications should not be rejected, it is unreasonable to expect a single rule-of-thumb cutoff for any (Δ)AFI to perform consistently across various models (Cheung and Lau, 2012; Pornprasertmanit et al., 2013).

Putnick and Bornstein (2016) found that 45.9% of studies they reviewed supplemented the LRT with at least one (Δ)AFI to draw conclusions about various levels of ME/I. Given the popularity of (Δ)AFIs, it is safe to assume any of those researchers who reported a significant LRT still did not reject their model if the (Δ)AFI(s) were within the guidelines of acceptable fit. The LRT appears to be used as the sole criterion to evaluate ME/I only half as often (16.7%) as (Δ)AFI(s) alone (34.1%), the most popular of which is the comparative fit index (CFI; Bentler, 1990), at least in the context of ME/I (Putnick and Bornstein, 2016). Given the sensitivity of (Δ)AFI sampling distributions to data and model characteristics (Marsh et al., 2004), basing conclusions about configural invariance on AFIs (e.g., interpreting CFI >0.95 as evidence of good approximate fit) leads to Type II errors in large samples, but can also lead to inflated Type I errors in small samples (Jorgensen et al., 2017). Permutation also provides a solution to problems with unknown (Δ)AFI sampling distributions by comparing observed configural-model AFIs to empirical sampling distributions derived under the H_0 of equivalent group configurations (Jorgensen et al., in press).

ILLUSTRATIVE EXAMPLE

To demonstrate the utility of the recently proposed permutation test and how multivariate modification indices can be used to modify a model under the assumption of configural invariance, I fit a three-factor multigroup CFA model with simple structure to the Holzinger and Swineford (1939) dataset, which has often been repurposed for illustrative examples (e.g., Jöreskog, 1969; Tucker and Lewis, 1973). A subset of the data are available as part of the `lavaan` package (Rosseel, 2012), including three indicators for each of three mental-ability constructs: visual, textual, and speed. This illustration assesses configural invariance across two schools (Pasteur: $N = 156$; Grant–White: $N = 145$), which is the most common number of groups analyzed (75%; Putnick and Bornstein, 2016). I provide R syntax for all analyses in the

TABLE 1 | Estimated parameters from CFA with simple structure.

Common factor	Indicator	Mental-Ability test description	Pasteur School		Grant–White School	
			λ	**θ**	**λ**	**θ**
Visual	X_1	Visual perception	1.047	0.298	0.777	0.715
	X_2	Cubes	0.412	1.334	0.572	0.899
	X_3	Lozenges	0.597	0.989	0.719	0.557
Textual	X_4	Paragraph comprehension	0.946	0.425	0.971	0.315
	X_5	Sentence completion	1.119	0.456	0.961	0.419
	X_6	Word meaning	0.827	0.290	0.935	0.406
Speed	X_7	Speeded addition	0.591	0.820	0.679	0.600
	X_8	Speeded counting of dots	0.665	0.510	0.833	0.401
	X_9	Speeded discrimination between straight and curved capital (uppercase) letters	0.545	0.680	0.719	0.535

λ, factor loading; θ, residual variance. Factor variances were fixed to 1. Saturated mean structure not presented. In the Pasteur school, visual–textual covariance = 0.484, visual–speed covariance = 0.299, and speed–textual covariance = 0.325. In the Grant–White school, visual–textual covariance = 0.541, visual–speed covariance = 0.523, and speed–textual covariance = 0.336. SEs not reported, but all parameters significantly differed from zero at α = 5%.

Appendix, and **Table 1** presents descriptions of indicators of each factor, as well as parameter estimates from the configural CFA model.

There is evidence that the configural model does not fit the data perfectly, $\chi^2_{(48)} = 115.85$, $p = 0.0000002$, and both CFI = 0.923 and RMSEA = 0.097, 90% CI [0.075, 0.120], suggest that the degree of misspecification is not ignorable, using Hu and Bentler's (1999) recommended cutoffs of CFA >0.95 and RMSEA < 0.06. Thus, the three-factor model with simple structure does not appear to adequately capture features of the data-generating process. Without additional information about group discrepancy, a researcher interested in modifying the model might begin by assessing model fit separately within each group. Similar results would be found for both the Pasteur school, $\chi^2_{(24)} = 64.31$, $p = 0.00002$, CFI = 0.903, RMSEA = 0.104, 90% CI [0.074, 0.135], and the Grant–White school, $\chi^2_{(24)} = 51.54$, $p = 0.001$, CFI = 0.941, RMSEA = 0.089, 90% CI [0.055, 0.122], leading to the conclusion that both groups' models require modification. But without informing the researcher about (lack of) evidence of group discrepancy, it would be unclear whether the most appropriate course of action would be to attempt freeing the same parameter(s) in both groups simultaneously or to modify each group's model independently.

Permutation Test

A permutation test of configural invariance can be conducted by comparing $\chi^2_{(48)} = 115.85$ to an empirical sampling distribution rather than a central χ^2 distribution with 48 *df*. An empirical sampling distribution under the H_0 of equivalent model configurations can be estimated by randomly reassigning rows of data to the two schools, fitting the configural model to the permuted data, and saving χ^2. Repeating these steps numerous times results in a permutation distribution of χ^2, and a *p* value can be calculated as the proportion of the distribution that exceeds (indicates worse fit than) the observed χ^2. Because the students are assumed equivalent when they are randomly reassigned to schools, the permutation distribution reflects the sampling variance of χ^2 under the assumption that the schools share the same data-generating model, but without assuming that the data-generating model corresponds perfectly with the fitted model. Due to poor model fit (i.e., the H_0 of no overall approximation discrepancy is rejected), the permutation distribution is not expected to approximate a central χ^2 distribution with 48 *df*, but it has been shown to approximate the sampling distribution under the H_0 of no group discrepancy (Jorgensen et al., 2017, in press). Likewise, CFI and RMSEA can be compared to permutation distributions, overcoming important limitations of AFIs: the lack of a theoretical sampling distribution for CFI, and the lack of consensus about a particular value of CFI or RMSEA that would indicate adequate approximate fit in all contexts.

A permutation test revealed no evidence against the H_0 of configural invariance using either χ^2($p = 0.19$), CFI ($p = 0.17$), or RMSEA ($p = 0.19$) as criterion. Thus, model modification can proceed by freeing the same parameter(s) in both groups simultaneously. This could minimize well documented problems with data-driven use of modification indices leading to models that do not generalize to new samples from the same population (MacCallum, 1986; MacCallum et al., 1992; French and Finch, 2008). The hypothesized CFA model fixes 18 cross-loadings and 36 residual covariances to zero in each of two groups, resulting in 108 modification indices for individual parameters (i.e., 1-*df* tests). Inspecting multivariate modification indices (i.e., 2-*df* tests) reduces the number of tests by half, from 108 to 54. More generally, with *g* groups, there will always be *g* times as many 1-*df* modification indices as *g*-*df* modification indices. Before presenting results for the CFA model, I elaborate further on the multivariate modification index.

Multivariate Modification Indices

My discussion below is in the context of maximum likelihood estimation, but the same concepts can be applied to other discrepancy functions for estimating SEM parameters (Bentler and Chou, 1992). Lagrange multipliers fit into a framework of three tests of parameter restrictions, including Wald tests and nested-model LRTs (Buse, 1982). The LRT requires fitting both a restricted (M_0) and unrestricted (M_1) model. The LRT statistic is calculated by comparing the log-likelihood (ℓ) of the data under each model: LRT $= -2 \times (\ell_0 - \ell_1)$. If the H_0 is true and distributional assumptions are met, the LRT statistic is asymptotically distributed as a central χ^2 random variable with *df* equal to the number of restrictions in M_0 relative to M_1.

The Wald and Lagrange multiplier tests are asymptotically equivalent to the LRT, but the Wald test only requires fitting M_1, whereas the Lagrange multiplier test only requires fitting M_0 (for details see Buse, 1982). The modification indices provided by most SEM software packages are 1-*df* Lagrange multipliers associated with each fixed parameter (or equality constraint), and they estimate the LRT statistic (i.e., the change in χ^2 of M_0) if that constraint were freed in M_1 (but without needing to fit M_1), assuming all other parameter estimates would remain unchanged between M_0 and M_1. Calculation of Lagrange multipliers utilizes information from the gradient (first derivative of the discrepancy function). Specifically, the curvature of the likelihood function evaluated with respect to the null-hypothesized value (θ_0) of a fixed parameter (typically zero) provides a clue about how far θ_0 is from the true θ, relative to the estimated sampling variability.

Bentler and Chou (1992) extended this simple idea to evaluating the curvature of the likelihood function in multiple dimensions with respect to a vector of constrained parameters. Multivariate Lagrange multipliers have only been implemented in some SEM software packages, such as EQS (Bentler, 2006) and PROC CALIS (SAS Institute Inc., 2011). In the spirit of the open-access *Frontiers* journal[4], my applied example utilizes the freely available open-source R package `lavaan` (Rosseel, 2012), which implements multivariate Lagrange multipliers via the `lavTestScore()` function, along with the widely available 1-*df* statistics via the `modificationIndices()` function. I discuss both in the context of the example CFA applied to the Holzinger and Swineford (1939) data set. As noted in previous research (e.g., MacCallum et al., 1992) and SEM textbooks (e.g., Brown, 2015; Kline, 2015), purely data-driven specification searches do not lead to generalizable, reproducible models, so model modifications should always be guided by substantive theory. The current study, however, is focused on the statistics themselves, so my interpretation of results focuses primarily on decisions that a hypothetical researcher might be influenced to make when inspecting modification indices.

Table 2 presents the largest 1-*df* modification indices from the CFA model with simple structure, six of which (three in each group) were above 10. These results do not provide unambiguous guidance about which parameter constraints should be released. The largest modification index is associated with a residual covariance between the seventh and eighth indicators (of the same factor) in the Grant–White group. The second largest modification index (very similar in value to the largest) is associated with a cross-loading of the ninth indicator (speeded discrimination between straight and curved letters) on the visual factor, also in the Grant–White group. This is also the only parameter that is significant for both groups, although it is not significant in the Pasteur group after a Bonferroni adjustment for multiple tests. Arguably, it may make theoretical sense to free this parameter given that the X_9 task required similar visual skills as the other visual indicators. If one considered the standardized expected parameter changes in tandem with modification indices, as advised by Saris et al. (2009) see also Whittaker (2012), then the cross-loading of the first indicator on the textual factor in the Pasteur group might be considered the best candidate instead.

TABLE 2 | Largest univariate and multivariate modification indices for fixed (to zero) parameters.

School	Parameter	MI	EPC	SEPC
Pasteur	Visual → X_9	11.07[a]	0.32	0.32
	Textual → X_1	10.18[a]	0.89	0.76
	$X_4 \longleftrightarrow X_6$	11.28[a]	−0.33	−0.29
Grant–White	Visual → X_7	11.27[a]	−0.39	−0.38
	Visual → X_9	24.54[a,b]	0.58	0.57
	$X_7 \longleftrightarrow X_8$	24.82[a,b]	0.61	0.57
Multivariate	Visual → X_7	16.45[a,b]		
(MI $= \hat{\chi}^2_{df=2}$)	Visual → X_9	35.61[a,b]		
	$X_7 \longleftrightarrow X_8$	29.01[a,b]		

MI, modification index. (S); EPC, (standardized) expected parameter change (unavailable for multivariate MIs). → indicates a factor loading. ⟷ indicates a covariance.
[a] Significant at α = 5%.
[b] Significant at Bonferroni-adjusted α =0.05/108 = 0.00046 (critical $\hat{\chi}^2_{df=1}$ = 12.26) for 1-df MIs, or α = 0.05/54 = 0.00093 (critical $\hat{\chi}^2_{df=2}$ = 10.97) for 2-df MIs.

The bottom rows of **Table 2** also present the significant 2-*df* modification indices, the largest of which was for the cross-loading of the ninth indicator on the visual factor, which was also the only parameter with a large 1-*df* modification index in both groups. The interpretation of these tests is less ambiguous because they formally test the same parameter constraint simultaneously in both groups, which the permutation test implied is appropriate because there is no evidence the group configurations differ. Freeing this parameter did lead to significantly better model fit, $\Delta\chi^2_{(2)} = 34.31$ (comparable to the expected $\chi^2 = 35.61$ in **Table 2**), $p = 0.00000004$, although the modified model still did not fit perfectly, $\chi^2_{(46)} = 81.55, p = 0.001$, CFI = 0.960, RMSEA = 0.072, 90% CI [0.045, 0.097]. Because the purpose of this application is merely to demonstrate tools for testing and modifying configural models, I do not consider further modifications of the example CFA.

Next, I present a small-scale simulation study designed to evaluate the use of multivariate modification indices. A concise simulation was designed to keep the focus on the purpose of this simulation, which is to provide a "proof of concept" that multivariate modification indices can control Type I errors better than univariate modification indices when the hypothesized

[4] As stated on the Frontiers web page (http://home.frontiersin.org/about/about-frontiers): "Our grand vision is to build an Open Science platform where everybody has equal opportunity to seek, share and generate knowledge, and that empowers researchers in their daily work."

model is approximately well specified but needs improvement. I focus on this situation because modification indices are unlikely to lead to the true data-generating model when a hypothesized model deviates substantially from it (MacCallum, 1986; MacCallum et al., 1992), and there is no reason to expect multivariate modification indices to perform differently in the latter situation.

METHODS

To simulate data in which the H_0 of configural invariance was true but the H_0 of exact model fit is false, I specified a two-factor CFA model for four groups, with three indicators for each of two common factors. The factor loadings were $\lambda = 0.6$, 0.7, and 0.8 for the first, second, and third indicator of each factor, respectively. The residual variances were specified as $1 - \lambda^2$ so that indicators were multivariate normal with unit variances. Factor variances were fixed at 1 (also in the analysis model, for identification), and all indicator and factor intercepts were zero. Factor correlations were 0.2, 0.3, 0.4, and 0.5 in Groups 1, 2, 3, and 4, respectively, so that population covariance matrices were not identical, although model configurations were equivalent.

Imperfect overall model fit was specified by setting two residual covariances in the four populations with values of 0.2 between the first and fourth indicators, corresponding to a moderate residual correlation of $0.2/0.64 = 0.31$, and 0.15 between the second and fifth indicators, corresponding to a moderate residual correlation of $0.15/0.51 = 0.29$. These parameters were specified in all groups, so the population models were configurally invariant. Fixing these two residual covariances to zero in the analysis model resulted in significant misfit, $\chi^2_{(32)} = 54.05, p = 0.009$, when the model was fit to the population covariance matrices, using samples sizes of $N = 100$ in each group. Approximate fit was questionable, acceptable CFI = 0.962, unacceptable RMSEA = 0.083, 90% CI [0.042, 0.120], so the configural model would have a considerable chance of being rejected when fit to a random sample drawn from this population. These fit measures are from the results of fitting the model to the population rather than sampled data, so they give an indication of the fit of the model, free from sampling error.

The configural model fixed six cross-loadings and 15 residual covariances to zero, yielding 21 modification indices to consider in each of four groups. The Bonferroni-adjusted α level was therefore $0.05/21 = 0.0024$ for 4-*df* simultaneous tests and $0.05/84 = 0.0006$ for 1-*df* tests; unadjusted α levels were not considered. I generated 1,000 random samples of $N = 100$ from each of the populations specified above, fit the configural model to the data, and recorded decisions about overall model fit (χ^2, CFI, and RMSEA) and model respecification (univariate and multivariate modification indices). Within each replication, I also used a permutation test of configural invariance. When the model needed respecification, the parameter with the largest significant 4-*df* modification index was freed in all groups, iteratively until no modification indices were significant. A replication was flagged for having made a familywise Type I error if in any iteration, the largest significant 4-*df* modification index belonged to any parameter besides the two omitted residual covariances; correct detections of the omitted parameters were also flagged to calculate power. Parameters were not freed on the basis of univariate modification indices, but I also recorded whether the largest significant 1-*df* modification index in the first iteration belonged to any parameter besides the two omitted residual covariances, as a basis for comparing the familywise Type I error rates of 4-*df* modification indices to a lower-bound for the familywise Type I error rates of 1-*df* modification indices.

RESULTS

Using overall model fit as the criterion for evaluating configural invariance led to rejecting the model in 99.9% of replications using a significant LRT as criterion. Using Hu and Bentler (1999) criterion for approximate model fit, the model was rejected in 93.9% of replications by CFI < 0.95 and 100% using RMSEA > 0.06. Thus, researchers using any of these criteria would frequently be motivated to modify their configural model. Knowing whether the data showed evidence of equivalent model configurations (despite poor fit) would therefore be very useful. The permutation test falsely rejected the H_0 of configural invariance in only 4.9% of the 1,000 replications, so the Type I error rate did not deviate substantially from the nominal $\alpha = 5\%$. This demonstration is consistent with previous results investigating the permutation method for testing ME/I in a two-group scenario (Jorgensen et al., 2017, in press). The unique contribution of this simulation, however, is to evaluate the performance of rarely utilized multivariate modification indices.

Multivariate modification indices correctly detected that at least one of the two omitted residual covariances should be freed in 99.6% of the replications, and correctly detected both omitted parameters in 73.9% of replications. This was accomplished while maintaining nominal (4.4%) familywise Type I errors across iterative modifications. By comparison, the largest 1-*df* modification index in the original configural model flagged an incorrect parameter in 9.5% of replications, implying that familywise Type I error rates would be at least that bad if they were instead used to iteratively modify the model. The poor performance of decisions based solely on 1-*df* modification indices is also consistent with previous results (MacCallum, 1986; MacCallum et al., 1992).

DISCUSSION

The aim of this paper was to advance two methods for testing configural invariance: how to test the correct H_0 and how to test constraints in a poor-fitting configural model. A recently developed tool is a permutation test of the H_0 of equivalent model configurations, which has shown promising control of Type I errors even when a configural model fits poorly (Jorgensen et al., 2017, in press). When the data show no strong evidence against the H_0, researchers might be motivated to explore ways to modify their model to better reflect the data-generating process. Multivariate

Lagrange multipliers (Bentler and Chou, 1992) can provide tests of constraints on the same parameter simultaneously across groups. A small-scale simulation illustrated how these could limit Type I errors better than traditional 1-*df* modification indices for individual fixed parameters within each group.

The simulation was not designed to provide comprehensive information across a variety of conditions, but it contributes some evidence that these tools warrant further investigation. Given that fully invariant metric (17.8%) and scalar (42.2%) models are rejected many times more often than configural (5.5%) models (Putnick and Bornstein, 2016), it is easier to find guidance in the literature about modifying metric and scalar models to establish partial invariance (e.g., Byrne et al., 1989; Vandenberg and Lance, 2000; Millsap, 2011). The current study therefore contributes to a sparser literature on modifying configural models, which Jorgensen et al. (2017, in press) showed might require more careful attention than common practice currently pays it. Note, however, that the current investigation does not address the issue of establishing "partial configural" invariance, but rather improving the fit of a configurally invariant model. More extensive investigations could shed light on the general applicability of the permutation test and of multivariate modification indices across a variety of conditions (e.g., different numbers of groups, sample sizes and ratios, varying other nonzero parameter values). For instance, the Holzinger and Swineford (1939) example application had only two groups, which may not be as prone to inflated Type I error rates as the four-group simulated data showed for 1-*df* modification indices.

This paper focused only on the situation when the H_0 of configural invariance was true. When the data provide evidence against the assumption of equivalent model configurations[5], more restrictive levels of invariance cannot be assumed either, nor would the proposed use of multivariate modification indices be relevant for modifying the model simultaneously across groups. If there are more than two groups, one could potentially test whether each pair of groups provide evidence against configural invariance, then test more restrictive levels of ME/I only for subsets that do not. Future research would be required to reveal whether Type I error rates could be maintained under such a follow-up procedure, but Jorgensen et al. (2017, in press) did find nominal error rates for the omnibus test of configural invariance with two-group data. According to Putnick and Bornstein (2016), most studies (75%) involve only two groups, so follow-up tests on subsets of groups might not be required often in practice.

I conclude by reiterating the importance of substantive theory to guide the process of model respecification (Brown, 2015; Kline, 2015). Purely data-driven use of modification indices tends to result in models that are over-fit to sample-specific nuances rather than mimicking the true data-generating process (MacCallum, 1986; MacCallum et al., 1992). Modification indices only tend to identify the correct parameter(s) to free when the model is already close to correctly specified, not when the model deviates substantially in form from the true model (MacCallum, 1986; MacCallum et al., 1992), so the same behavior should be expected from the multivariate modification indices applied to simultaneous changes in a single model across groups. Assuming the configural model is close to correctly specified, expected parameter changes may also provide useful supplementary information to use in tandem with modification indices (Saris et al., 2009; Whittaker, 2012), but like modification indices, their validity rests on the assumption that the structure of the model is basically correct except that at least one parameter constraint is not near its true population value. Hayduk (2014) showed that this may not be a safe assumption, given that factor models can fit data patterns from very different kinds of models, so poorly fitting factor models might be misspecified in ways beyond fixing too many parameters to zero. Correlation residuals provide information about model inadequacy in terms of the data pattern that the model tries to reproduce, so their inspection might be more likely to help a researcher speculate about different kinds of data-generating models. However, Lagrange multipliers are useful for testing specific hypotheses about parameter constraints, which are asymptotically equivalent to a LRT but only require fitting the constrained model rather than many less restricted models.

AUTHOR CONTRIBUTIONS

TJ is responsible for the data analysis (using openly available data), design the simulation study, and writing the manuscript.

[5] See Jorgensen et al. (2017; in press) for an investigation of power to detect different model configurations.

REFERENCES

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychol. Bull.* 107, 238–246. doi: 10.1037/0033-2909.107.2.238

Bentler, P. M. (2006). *EQS 6 Structural Equations Program Manual*. Encino, CA: Multivariate Software, Inc.

Bentler, P. M., and Chou, C.-P. (1992). Some new covariance structure model improvement statistics. *Sociol. Methods Res.* 21, 259–282. doi: 10.1177/0049124192021002006

Brown, T. A. (2015). *Confirmatory Factor Analysis for Applied Research,* 2nd Edn. New York, NY: Guilford.

Buse, A. (1982). The likelihood ratio, wald, and lagrange multiplier tests: an expository note. *Am. Stat.* 36, 153–157. doi: 10.2307/2683166

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

Cheung, G. W., and Lau, R. S. (2012). A direct comparison approach for testing measurement invariance. *Organ. Res. Methods* 15, 167–198. doi: 10.1177/1094428111421987

Cheung, G. W., and Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Struct. Equation Model.* 9, 233–255. doi: 10.1207/S15328007SEM0902_5

Cudeck, R., and Henly, S. J. (1991). Model selection in covariance structures analysis and the "problem" of sample size: a clarification. *Psychol. Bull.* 109, 512–519. doi: 10.1037/0033-2909.109.3.512

French, B. F., and Finch, W. H. (2008). Multigroup confirmatory factor analysis: locating the invariant referent sets. *Struct. Equation Model.* 15, 96–113. doi: 10.1080/10705510701758349

Hayduk, L. (2014). Seeing perfectly fitting factor models that are causally misspecified: understanding that close-fitting models can be worse. *Educ. Psychol. Meas.* 74, 905–926. doi: 10.1177/0013164414527449

Holzinger, K., and Swineford, F. (1939). *A Study in Factor Analysis: the Stability of a Bifactor Solution. Supplementary Educational Monograph, no. 48*. Chicago, IL: University of Chicago Press.

Hu, L.-t., and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struct. Equation Model.* 6, 1–55. doi: 10.1080/10705519909540118

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

Jorgensen, T. D., Kite, B. A., Chen, P.-Y., and Short, S. D. (in press). Permutation randomization methods for testing measurement equivalence and detecting differential item functioning in multiple-group confirmatory factor analysis. *Psychol. Methods*. doi: 10.1037/met0000152

Jorgensen, T. D., Kite, B., Chen, P.-Y., and Short, S. D. (2017). "Finally! A valid test of configural invariance using permutation in multigroup CFA," in *Quantitative Psychology: the 81st Annual Meeting of the Psychometric Society, Asheville, North Carolina (2016),* eds L. A. van der Ark, M. Wiberg, S. A. Culpepper, J. A. Douglas, and W.-C. Wang (New York, NY: Springer), 93–103.

Kline, T. A. (2015). *Principles and Practice of Structural Equation Modeling*, 4th Edn. New York, NY: Guilford.

MacCallum, R. C. (1986). Specification searchers in covariance structure modeling. *Psychol. Bull.* 100, 107–120. doi: 10.1037/0033-2909.100.1.107

MacCallum, R. C. (2003). 2001 presidential address: working with imperfect models. *Multivariate Behav. Res.* 38, 113–139. doi: 10.1207/S15327906MBR3801_5

MacCallum, R. C., Roznowski, M., and Necowitz, L. B. (1992). Model modifications in covariance structure analysis: the problem of capitalization on chance. *Psychol. Bull.* 111, 490–504. doi: 10.1037/0033-2909.111.3.490

Marsh, H. W., Hau, K.-T., and Wen, Z. (2004). In search of golden rules: comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Struct. Equation Model.* 11, 320–341. doi: 10.1207/s15328007sem1103_2

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika* 58, 525–543. doi: 10.1007/BF02294825

Millsap, R. E. (2011). *Statistical Approaches to Measurement Invariance*. New York, NY: Routledge.

Nevitt, J., and Hancock, G. R. (2004). Evaluating small sample approaches for model test statistics in structural equation modeling. *Multivariate Behav. Res.* 39, 439–478. doi: 10.1207/S15327906MBR3903_3

Pornprasertmanit, S., Wu, W., and Little, T. D. (2013). "Using a Monte Carlo approach for nested model comparisons in structural equation modeling," in *New Developments in Quantitative Psychology,* eds R. E. Millsap, L. A. van der Ark, D. M. Bolt, and C. M. Woods (New York, NY: Springer), 187–197.

Putnick, D. L., and Bornstein, M. H. (2016). Measurement invariance conventions and reporting: the state of the art and future directions for psychological research. *Dev. Rev.* 41, 71–90. doi: 10.1016/j.dr.0.2016.06.004

R Core Team (2017). *R: A Language and Environment for Statistical Computing (Version 3.3.3) [Computer software]*. Vienna: R Foundation for Statistical Computing. Available online at: https://www.R-project.org/

Reise, S. P., Widamin, K. F., and Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. *Psychol. Bull.* 114, 552–566. doi: 10.1037/0033-2909.114.3.552

Rosseel, Y. (2012). `lavaan`: an R package for structural equation modeling. *J. Stat. Softw.* 48, 1–36. doi: 10.18637/jss.v048.i02

Saris, W. E., Satorra, A., and van der Veld, W. M. (2009). Test structural equation models or detection of misspecifications? *Struct. Equat. Model.* 16, 561–582. doi: 10.1080/10705510903203433

SAS Institute Inc. (2011). *SAS/STAT® 9.3 User's Guide*. Cary, NC: Author.

Steiger, J. H., and Lind, J. C. (1980). "Statistically-Based Tests for the Number of Common Factors," in *Paper Presented at the Annual Meeting of the Psychometric Society* (Iowa City, IA).

Tucker, L. R., and Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika* 38, 1–10. doi: 10.1007/BF02291170

Vandenberg, R. J., and Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. *Organ. Res. Methods* 3, 4–69. doi: 10.1177/109442810031002

Whittaker, T. A. (2012). Using the modification index and standardized expected parameter change for model modification. *J. Exp. Educ.* 80, 26–44. doi: 10.1080/00220973.2010.531299

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

APPENDIX

R Syntax for Applied Example

```
  ## use data available in lavaan package
  library(lavaan)
  HS <- lavaan::HolzingerSwineford1939
  ## specify configural invariance model
  mod.config <- ’
  visual =~ x1 + x2 + x3
  textual =~ x4 + x5 + x6
  speed =~ x7 + x8 + x9
  ’
  ## fit model to schools, print results
  fit.config <- cfa(mod.config, data = HS, std.lv = TRUE, group = "school")
  summary(fit.config, fit = TRUE)
  fitMeasures(fit.config, c("chisq","df","pvalue","cfi","rmsea","rmsea.ci.lower",
  "rmsea.ci.upper"))
  ## fit model separately per school
  fit.Pasteur <- cfa(mod.config, data = HS[HS$school == "Pasteur",], std.lv = TRUE)
  fitMeasures(fit.Pasteur, c("chisq","df","pvalue","cfi","rmsea",
  "rmsea.ci.lower","rmsea.ci.upper"))
  fit.Grant <- cfa(mod.config, data = HS[HS$school == "Grant-White",], std.lv = TRUE)
  fitMeasures(fit.Grant, c("chisq","df","pvalue","cfi","rmsea",
  "rmsea.ci.lower","rmsea.ci.upper"))
  ## Permutation Test using lavaanList()
  set.seed(3141593)
  dataList <- lapply(1:200, function(i) {HS$school <- sample(HS$school); HS})
  out.site <- cfaList(mod.config, dataList = dataList, std.lv = TRUE,
store.slots = NULL, group = "school", FUN = function(x) lavaan::fitMeasures(x,
c("chisq","cfi","rmsea")), parallel = "snow", ncpus = 3, iseed = 3141593)
  PF <- as.data.frame(do.call(rbind, out.site@funList))
  OF <- fitMeasures(fit.config, c("chisq","cfi","rmsea"))
  mean(PF[["chisq"]] > OF["chisq"])
  mean(PF[["cfi"]] < OF["cfi"])
  mean(PF[["rmsea"]] > OF["rmsea"])
  ## Permutation Test also available in the semTools package
  # library(semTools)
  # permuteMeasEq(nPermute = 200, con = fit.config, AFIs = c("chisq","cfi","rmsea"))
  ## inspect univariate (1-df) and multivariate (2-df) modification indices
  MI1 <- modindices(fit.config)
  MI1$p.value <- pchisq(MI1$mi, df = 1, lower.tail = FALSE)
  MI1$bonf <- p.adjust(MI1$p.value, method = "bonferroni")
  MI1[MI1$mi > 10,]
  MI1[MI1$bonf <0.05,]
  ## multivariate tests require changing the lavTestScore() source code in lavaan.
  ## Source code for the myScoreTest() function is available from the author on request.
  MI2 <- do.call(rbind, lapply(unique(paste0(MI1$lhs, MI1$op, MI1$rhs)), function(x)
{out <- myScoreTest(fit.config, add = x, univariate = FALSE)$test out$test <- x out}))
  MI2$bonf <- p.adjust(MI2$p.value, method = "bonferroni")
  MI2[MI2$bonf <0.05,]
  ## Fit model with cross-loading freed
  fit.cross <- cfa(c(mod.config, ’visual =~ x9’), data = HS, std.lv = TRUE, group =
"school")
  summary(fit.cross, fit = TRUE)
  anova(fit.config, fit.cross)
```

frontiers
in Psychology

METHODS
published: 24 October 2017
doi: 10.3389/fpsyg.2017.01823

Advances in Measurement Invariance and Mean Comparison of Latent Variables: Equivalence Testing and A Projection-Based Approach

*Ge Jiang, Yujiao Mai and Ke-Hai Yuan**

Department of Psychology, University of Notre Dame, Notre Dame, IN, United States

Measurement invariance (MI) entails that measurements in different groups are comparable, and is a logical prerequisite when studying difference or change across groups. MI is commonly evaluated using multi-group structural equation modeling through a sequence of chi-square and chi-square-difference tests. However, under the conventional null hypothesis testing (NHT) one can never be confident enough to claim MI even when all test statistics are not significant. Equivalence testing (ET) has been recently proposed to replace NHT for studying MI. ET informs researchers a size of possible misspecification and allows them to claim that measurements are practically equivalent across groups if the size of misspecification is smaller than a tolerable value. Another recent advancement in studying MI is a projection-based method under which testing the cross-group equality of means of latent traits does not require the intercepts equal across groups. The purpose of this article is to introduce the key ideas of the two advancements in MI and present a newly developed R package `equaltestMI` for researchers to easily apply the two methods. A real data example is provided to illustrate the use of the package. It is advocated that researchers should always consider using the two methods whenever MI needs to be examined.

Keywords: equivalence testing, measurement invariance, minimum tolerable size, projection method, scalar invariance

Edited by:
Oi-Man Kwok,
Texas A&M University, United States

Reviewed by:
Mark Lai,
University of Cincinnati, United States
Eun Sook Kim,
University of South Florida,
United States

****Correspondence:***
Ke-Hai Yuan
kyuan@nd.edu

Specialty section:
This article was submitted to
Quantitative Psychology and
Measurement,
a section of the journal
Frontiers in Psychology

Received: *02 August 2017*
Accepted: *02 October 2017*
Published: *24 October 2017*

Citation:
Jiang G, Mai Y and Yuan K-H (2017)
Advances in Measurement Invariance
and Mean Comparison of Latent
Variables: Equivalence Testing and A
Projection-Based Approach.
Front. Psychol. 8:1823.
doi: 10.3389/fpsyg.2017.01823

1. INTRODUCTION

Reliable and valid measurements are key to social and behavioral sciences. When studying difference across groups, an equally important concept is measurement invariance (MI) or equivalence (Mellenbergh, 1989; Meredith, 1993; Millsap, 2011; Kim et al., 2012), which entails that measurements in different groups are comparable. Equivalent measurements are logical prerequisites to the evaluation of substantive hypotheses, regardless of whether the interest is as simple as a test of mean difference between groups or as complex as a test for possible changes of theoretical constructs across groups (Vandenberg and Lance, 2000). In particular, the observed or estimated cross-group difference can be simply due to different types of attributes being measured across populations, rather than the difference in the same attribute. Then, the observed cross-group difference is not interpretable nor valid for quantifying the cross-group difference on the target attribute.

The most widely used approach to examine MI is multi-group structural equation modeling (SEM) which relies on a sequence of chi-square and chi-square-difference tests (Sörbom, 1974; Horn et al., 1983; Meredith, 1993). With multi-group SEM, the test of MI typically starts with the equality of population covariance matrices across groups. Rejection of this equality does not imply that the groups are not comparable. A series of tests are then conducted to identify the source of non-equivalence (e.g., factor structure, factor loadings, etc.) and also to determine the degree of equivalence. Equality constraints are added in a logical order, and the models being tested also become increasingly more restrictive (Byrne, 2010). Two models are connected by each set of equality constraints: a base model and a nested (constrained) model. The normal-distribution-based maximum likelihood (NML) is typically used to estimate the models, and we also have a test statistic that approximately follows a chi-square distribution. The difference between the values of the test statistic at the base and restricted models is called the chi-square-difference statistic, which is commonly used to evaluate the plausibility of the constraints. The most widely used statistic is the likelihood ratio test statistic corresponding to NML estimation, which is also what we use in this article.

Two major concerns exist over the multi-group SEM approach to MI. First, there is a logical issue when the conventional null hypothesis testing (NHT) is used to establish equivalence of measurements. In every step of MI tests, whenever the chi-square or chi-square-difference statistic is not significant at a given level (e.g., $\alpha = 0.05$), we move to the next step of the analysis by assuming that the current model under the null hypothesis holds. However, a non-significant test statistic does not imply that the involved model is correct or the involved components are invariant across groups. This is because NHT is constructed to reject the null hypothesis, and one can never be confident enough to claim equivalence even when all the statistics are not significant. Under such a practice, any violation against the previous hypotheses will be carried over to the next test. Yuan and Chan (2016) contains an example in which a sequence of tests for endorsing MI yields a rather different conclusion from that of testing the equality of covariance matrices across groups. Yuan and Bentler (2004) also showed that nested chi-square test is unable to control type I errors when the based model is misspecified, and the power of the test can also become rather weak.

Second, multi-group SEM approach for MI requires the intercepts of the manifest variables to be equal across groups before the means of latent constructs can be estimated (Sörbom, 1974). The cross-group equality of intercepts is commonly called *scalar invariance* (Horn and McArdle, 1992). The review by Vandenberg and Lance (2000) indicated that scalar invariance is rarely satisfied in practice. Marsh et al. (2017) also noted that "scalar invariance is an unachievable ideal that in practice can only be approximated." However, without scalar invariance, the means of the latent constructs cannot be compared under the conventional setup. Such a requirement greatly limits the use of the multi-group SEM approach to mean comparison of latent variables.

To address the first issue regarding the use of NHT, Yuan and Chan (2016) recently proposed using equivalence testing (ET) to replace NHT in multi-group SEM. In a sequence of tests for MI under ET, researchers are informed about a possible misspecification in every step, which enables them to effectively control the size of misspecification. Researchers can evaluate their results based on their own degrees of tolerance or using the adjusted cutoff values in connection with established rules of labeling the goodness of model fit in SEM. Yuan and Chan (2016) illustrated their approach using a simulated example with 2 groups, 9 variables, and 3 latent factors. They also provided an R program to compute the minimum tolerable size and adjusted cutoff values of fit indices for evaluating the goodness of the model under ET. However, one has to use a separate program to estimate the SEM model under different constraints before conducting ET using their R program. Thus, it is rather difficult for substantive researchers to correctly perform or interpret results at each step of the sequence of conducting the tests for MI. Also, although ET has been used in many areas of psychological and educational research, there is no self-contained software for conducting ET using chi-square and chi-square difference tests, especially for the purpose of MI. Our experience indicates that a statistical package must be in place before any new cutting-edge methodology can be applied by substantive researchers. Thus, we have developed an all-in-one R package `equaltestMI` that will be introduced in this article. Our illustration of the package with real data will also contribute to promoting ET in substantive areas where MI is routinely used in group comparison.

To address the second issue with the multi-group SEM approach to MI, Deng and Yuan (2016) proposed a new projection method to circumvent the scalar-invariance assumption by decomposing the observed means of the manifest variables into two orthogonal components. One component represents the means of the common scores and the other represents the mean of the specific factors. These two components are uniquely identified although the means of specific factors have been ignored in conventional factor analysis (Harman, 1976; Gorsuch, 1983). As we will see, the projection method allows us to test the cross-group equality of latent means independently from that of specific factors, and there is no need to constrain intercepts to be equal in this approach. In particular, only factor loadings are required equal across groups for conducting mean comparison of latent constructs. However, Deng and Yuan (2016) only presented the projection method using conventional NHT, not ET. Thus, the method still has the logical issues inherited from NHT, which will be addressed in this article by putting the projection method under ET.

The contributions of the current article are as follows: (1) using plain language to introduce the key ideas of ET and the projection method for examining MI; (2) combine ET and the projection method to provide valid inference on the tests of equality of latent factors and specific factors; (3) developing an accompanying R package `equaltestMI` so that substantive researchers can easily apply the two new methods as well as combining them in conducting MI analysis; and (4) providing a detailed tutorial to illustrate the use of `equaltestMI` with a real data example.

In the following sections, we first briefly review the types of tests in the conventional approach to MI. Then, we introduce the ET framework and the projection method. Next, we provide a step-by-step tutorial to illustrate the use of the accompanying R package `equaltestMI` with a real data example. We conclude this article with some remarks on the two new methods and the use of the R package.

2. METHODS

This section introduces two recent methodological advancements in examining MI. By avoiding the logical problem and unrealistic assumptions with the conventional multi-group SEM approach, the new methods provide a more valid platform for studying MI. In particular, ET is proposed to replace the NHT framework and the projection method is proposed to replace the tests of mean structure under the conventional multi-group SEM as developed in Sörbom (1974). To help introduce the two new methods, we first review the models and notations used in multi-group SEM and the sequence of tests for examining MI.

2.1. Multi-Group SEM

Suppose a set of p variables are collected for each of m groups, and they are obtained by administering the same instrument or properly adjusted to be on the same scale. Let $\mathbf{x}^{(j)}$ represent the vector of variables in the population for group j, $j = 1, .., m$, and the following SEM model holds within each group:

$$\mathbf{x}^{(j)} = \boldsymbol{\gamma}^{(j)} + \Lambda^{(j)}\boldsymbol{\xi}^{(j)} + \boldsymbol{\varepsilon}^{(j)},\ j = 1, \cdots, m, \qquad (1)$$

where the superscript (j) indicates the group membership; $\boldsymbol{\gamma}^{(j)}$ is a vector of p intercepts of the manifest variables, $\Lambda^{(j)}$ is $p \times k$ matrix of factor loadings, $\boldsymbol{\xi}^{(j)}$ is a vector of k factor scores, and $\boldsymbol{\varepsilon}^{(j)}$ is a vector of p errors. We assume that errors are uncorrelated and $\Psi^{(j)} = \text{Cov}(\boldsymbol{\varepsilon}^{(j)})$ is a diagonal matrix. The errors and the factors are also assumed to be uncorrelated with $E(\boldsymbol{\xi}^{(j)}) = \boldsymbol{\tau}^{(j)}$ and $\text{Cov}(\boldsymbol{\xi}^{(j)}) = \Phi^{(j)}$. It follows from Equation (1) that the model-implied mean and covariance structures for the m groups are respectively

$$\begin{aligned}\boldsymbol{\mu}^{(j)} &= \boldsymbol{\gamma}^{(j)} + \Lambda^{(j)}\boldsymbol{\tau}^{(j)} \text{ and}\\ \Sigma^{(j)} &= \Lambda^{(j)}\Phi^{(j)}\Lambda^{'(j)} + \Psi^{(j)},\ j = 1, \cdots, m.\end{aligned} \qquad (2)$$

Note that different groups might have different structures in (2), and $\boldsymbol{\gamma}^{(j)}$, $\Lambda^{(j)}$, $\boldsymbol{\tau}^{(j)}$, $\Phi^{(j)}$, and $\Psi^{(j)}$ are free to vary. The key point here is that the latent variables $\xi^{(j)}$ cannot be directly observed and must be measured with a set of manifest variables. These are standard assumptions in structural equation modeling and factor analysis, not particular to MI. With these notations, the steps of tests of MI (Vandenberg and Lance, 2000) and their corresponding chi-square and chi-square-difference statistics are given in **Table 1**. In the table, each subscript of the letter H represents the hypothesis for the involved parameters; and the subscripts of T represent the joint hypotheses under which the statistic is computed; while the superscript of T represents the hypothesis being tested by the nested chi-square statistic.

TABLE 1 | Types and steps of tests with the conventional approach to measurement invariance.

Step	Hypothesis	Name	Test statistics	
			Overall model	Nested model
1	$H_\sigma : \Sigma^{(1)} = \cdots = \Sigma^{(m)}$		T_σ	
2	$H_C : \Sigma^{(j)} = \Sigma(\boldsymbol{\theta}^{(j)})$	configural	T_C	
3	$H_\lambda : \Lambda^{(1)} = \cdots = \Lambda^{(m)}$	metric	$T_{C\lambda}$	$T_C^\lambda = T_{C\lambda} - T_C$
4a	$H_\psi : \Psi^{(1)} = \cdots = \Psi^{(m)}$		$T_{C\lambda\psi}$	$T_{C\lambda}^\psi = T_{C\lambda\psi} - T_{C\lambda}$
5a	$H_\phi : \Phi^{(1)} = \cdots = \Phi^{(m)}$		$T_{C\lambda\psi\phi}$	$T_{C\lambda\psi}^\phi = T_{C\lambda\psi\phi} - T_{C\lambda\psi}$
4b	$H_\gamma : \boldsymbol{\gamma}^{(1)} = \cdots = \boldsymbol{\gamma}^{(m)}$	scalar	$T_{C\lambda\gamma}$	$T_{C\lambda}^\gamma = T_{C\lambda\gamma} - T_{C\lambda}$
5b	$H_\tau : \boldsymbol{\tau}^{(1)} = \cdots = \boldsymbol{\tau}^{(m)}$		$T_{C\lambda\gamma\tau}$	$T_{C\lambda\gamma}^\tau = T_{C\lambda\gamma\tau} - T_{C\lambda\gamma}$
4c	$H_\gamma : \boldsymbol{\gamma}^{(1)} = \cdots = \boldsymbol{\gamma}^{(m)}$	scalar	$T_{C\lambda\gamma}$	$T_{C\lambda}^\gamma = T_{C\lambda\gamma} - T_{C\lambda}$
5c	$H_\psi : \Psi^{(1)} = \cdots = \Psi^{(m)}$		$T_{C\lambda\gamma\psi}$	$T_{C\lambda\gamma}^\psi = T_{C\lambda\gamma\psi} - T_{C\lambda\gamma}$
6c	$H_\tau : \boldsymbol{\tau}^{(1)} = \cdots = \boldsymbol{\tau}^{(m)}$		$T_{C\lambda\gamma\psi\tau}$	$T_{C\lambda\gamma\psi}^\tau = T_{C\lambda\gamma\psi\tau} - T_{C\lambda\gamma\psi}$

Following the work of Sörbom (1974) and Jöreskog (1971), the tests of MI usually start with a test of equality of the population covariance matrices. Statistically speaking, the first step tests H_σ : $\Sigma^{(1)} = \cdots = \Sigma^{(m)}$, where $\Sigma^{(j)}$ is the population covariance matrix of group j. A non-significant statistic of this test is generally regarded as an endorsement of overall measurement equivalence. However, a significant test statistic does not mean that the involved groups are not comparable and it is necessary to conduct subsequent tests to identify the sources of non-equivalence. To test if any aspects of the groups are invariant, a common SEM model is assumed and the equalities of its components across groups are tested in an increasingly restrictive fashion. In step 2, the SEM model is fitted to each group separately and one examines if the same model structure holds across groups (configural invariance). We denote configural invariance as H_c : $\Sigma^{(j)} = \Sigma(\boldsymbol{\theta}^{(j)})$, $j = 1, \cdots, m$, implying that the same structured model $\Sigma(\boldsymbol{\theta}^{(j)})$ holds in all the groups but their parameters $\boldsymbol{\theta}^{(j)}$ can differ across groups. If H_c holds, configural invariance is established, and one tests the equality of factor loading matrices (metric invariance) in step 3. We denote metric invariance as H_λ : $\Lambda^{(1)} = \cdots = \Lambda^{(m)}$, implying that the factor loadings are invariant across all the groups. After both configural (H_c) and metric invariances (H_λ) are established, one next separately tests the equalities in covariance structure and mean structure. For the covariance structure, one first tests the equality of error variances $\Psi^{(j)}$ across groups; and if that holds, one then tests the equality of factor covariance matrices $\Phi^{(j)}$ across groups.

For the mean structure, two types of invariance have been conceptualized (Meredith, 1993; Vandenberg and Lance, 2000). Measurements satisfying Steps 2, 3, and 4b are called *strong invariance* (Meredith, 1993), while those satisfying Steps 2, 3, 4c, and 5c are called *strict invariance*. For either of the invariances, the equality of intercepts of manifest variables (scalar invariance, H_γ) is tested first. If scalar invariance holds, strong invariance is achieved, and one continues to test the equality of latent means (H_τ). To achieve strict invariance, one needs to test the equality of error variances $\Psi^{(j)}$ after scalar invariance and then tests the equality of latent means. In summary, the steps to examine MI of covariance structure is $1 \rightarrow 2 \rightarrow 3 \rightarrow 4a \rightarrow 5a$, the sequence for testing mean structure and achieving strong invariance is $1 \rightarrow 2 \rightarrow 3 \rightarrow 4b$, and the sequence for achieving strict invariance is $1 \rightarrow 2 \rightarrow 3 \rightarrow 4c \rightarrow 5c$. Step 5b or 6c might not be needed if the interest of the MI analysis is to compare individuals. But the test of H_τ will be the ultimate goal if the interest is to compare groups, as in ANOVA or *t*-test.

To test the hypotheses mentioned above, chi-square and chi-square-difference statistics are computed in the last two columns of **Table 1** with the superscripts and subscripts denoting the involved hypotheses. In any of the three sequences above, a model with more constraints is nested in a model with fewer constraints and the additional constraints can be tested using the chi-square-difference statistic. For example, the statistic $T_{c\lambda\gamma}$ in step 4b of **Table 1** evaluates the joint hypothesis $H_{c\lambda\gamma} = H_c + H_\lambda + H_\gamma$, and the model under $H_{c\lambda\gamma}$ is nested in the model under $H_{c\lambda}$. The corresponding chi-square-difference statistic $T_{c\lambda}^{\gamma}$ evaluates the additional constraints under H_γ, and it is computed as the difference between $T_{c\lambda\gamma}$ and $T_{c\lambda}$, i.e., $T_{c\lambda}^{\gamma} = T_{c\lambda\gamma} - T_{c\lambda}$.

2.2. Equivalence Testing

ET was proposed to address the logical issues with NHT to establish equivalence of measures across groups (Yuan and Chan, 2016). A major distinction between ET and NHT is the formulation of null hypothesis. The null hypothesis under NHT is that the model or constraints hold in the population, whereas the null hypothesis under ET is that the size of misspecification in the model or constraints is greater than a tolerable value. When the null hypothesis is rejected under ET at level α, we are confident with probability $1-\alpha$ that the size of misspecification is less than or equal to the tolerable value. Consequently, the current model or components are deemed[1] as MI, and we continue with testing the subsequent hypothesis. Otherwise, we declare that the size of misspecification in the current model or hypothesis is not tolerable and stop at the previous level of equivalence. We will further discuss the specification of tolerable values using the fit index RMSEA (root mean square error of approximation, Steiger and Lind, 1980).

As with NHT, we need to have a statistic to work with under ET. In this article, we use the likelihood ratio statistic $T_{ml} = (N-m)F_{ml}$, where N is the total sample size across the m groups and F_{ml} is the normal-distribution-based discrepancy function proportionally weighted according to the sample sizes in the m groups (e.g., Equations 23 and 4 in Yuan and Bentler, 2006). Let F_{ml0} be the population counterpart of F_{ml}, the null hypothesis under NHT is $H_0 : F_{ml0} = 0$ whereas that under ET is

$$H_{e0} : F_{ml0} > \epsilon_0 \tag{3}$$

with ϵ_0 being a small positive number that one can tolerate for the size of misspecification. As for NHT, we need to assume that T_{ml} follows a central chi-square distribution χ^2_{df} when $F_{ml0} = 0$ and a non-central chi-square distribution $\chi^2_{df}(\delta)$ when $F_{ml0} > 0$, where $\delta = (N-m)F_{ml0}$ is the non-centrality parameter (ncp). Let $\delta_0 = (N-m)\epsilon_0$ and $c_\alpha(\delta_0)$ be the left-tail critical value of $\chi^2_{df}(\delta_0)$ at level α. Then we reject the null hypothesis H_{e0} in (3) when $T_{ml} < c_\alpha(\delta_0)$ and the type I error is controlled at level α. When the H_{e0} in (3) is rejected, we conclude that the size of misspecification of the current model is no greater than ϵ_0 with $1-\alpha$ confidence.

Similarly, when the chi-square-difference statistic is formulated according to T_{ml}, and our null hypothesis under ET is

$$H_{eab} : F_{mla0} - F_{mlb0} > \epsilon_{0ab}, \tag{4}$$

where ϵ_{0ab} is a tolerable value of misspecification due to the additional constraints in model *A* beyond that in the based model *B*. When the difference statistic is smaller than the left-tail critical value corresponding to $\chi^2_{dfab}(\delta_{0ab})$ with $\delta_{0ab} = (N-m)\epsilon_{0ab}$, we reject H_{eab} and conclude with probability $1-\alpha$ that the size of misspecification due to the additional constraints in model *A* (beyond that in model *B*) is smaller than the tolerable value or is tolerable.

The specification of a tolerable value ϵ_0 is crucial for ET. Although any choice of ϵ_0 cannot avoid an arbitrary nature, it is a necessary element for conducting ET. Following Yuan and Chan (2016), we specify ϵ_0 by relating it to the population value of the fit index RMSEA through

$$\epsilon_0 = df(\text{RMSEA}_0)^2/m, \tag{5}$$

where $\text{RMSEA}_0 = \{m\delta_0/[df(N-m)]\}^{1/2}$ with δ_0 being the ncp of the nominal chi-square distribution corresponding to the statistic T_{ml} with m groups (Steiger, 1998). With respect to the use of the conventional[2] RMSEA, MacCallum et al. (1996) suggested cutoff[3] values 0.01, 0.05, 0.008, and 0.10 to distinguish between excellent, close, fair, mediocre, and poor fit, respectively. As can be seen from Equation (5), when other terms are held constant, the larger the value of ϵ_0, the larger the RMSEA_0 is. This means that for a given model, a larger tolerable value of misspecification ϵ_0 implies that we allow for a less ideal model as quantified by RMSEA_0. There are two ways to use the relationship in Equation (5) to evaluate the fit of the current model. One can obtain the values of ϵ_0 corresponding to $\text{RMSEA}_0 = 0.01$, 0.05, 0.08, and 0.10, respectively, and compare T_{ml} against the critical values $c_\alpha(\epsilon_0)$ with those ϵ_0 values. If T_{ml} is between the critical values corresponding to $\text{RMSEA}_0 = 0.01$ and 0.05, then the model achieves a close fit for the observed samples.

[1] Even if MI does not hold literally, the violation against MI is small enough that we can comfortably ignore it.

[2] The value of the conventional RMSEA is computed according to the value of the observed test statistic T_{ml} whereas RMSEA_0 in Equation (5) is related to the value of ϵ_0 in Equation (3) or ϵ_{0ab} in Equation (4), and is used for the purpose of ET.

[3] These cutoff values are necessary for labeling the goodness of model fit but may be of limited scientific value (see Lai and Green, 2016).

Alternatively, we can solve the equation

$$T_{ml} = c_\alpha(\epsilon_t). \tag{6}$$

for the value of ϵ_t, which is an increasing function of T_{ml}. Unlike the ϵ_0 in Equation (3) or (5) that is specified a priori, the ϵ_t in (6) is data dependent. However, rejection of the hypothesis in (3) is equivalent to $\epsilon_t < \epsilon_0$. Yuan and Chan (2016) called the ϵ_t in (6) *the minimum tolerable size* (T-size) of misspecification. If one cannot tolerate the T-size ϵ_t, then hypothesis with any prespecified ϵ_0 that is less than ϵ_t cannot be rejected since $T_{ml} > c_\alpha(\epsilon_0)$, and we will not be able to continue with the analysis in the sequence of endorsing MI. Let RMSEA$_t$ be the value of RMSEA defined at ϵ_t. ET can be equivalently conducted using the established cutoff values of RMSEA and the values of RMSEA$_t$ corresponding to the ϵ_t in (6). We will illustrate this procedure in a later section via a real data example.

Compared with the conventional methods, ET informs us the size of a possible misspecification at each step of endorsing MI, and it is still up to the researcher to decide whether the size is tolerable. Established values of RMSEA facilitate us to make a decision on the size of misspecification according to the values of RMSEA$_t$. However, the conventional cutoff values of RMSEA are too stringent to evaluate the model fit under ET, and the cutoff values need to be modified accordingly. Yuan and Chan (2016) developed formulas of adjusted cutoff values for evaluating RMSEA$_t$ so that labeling of goodness of fit is comparable to evaluating the conventional RMSEA by existing cutoff values. These formulas are incorporated in our R package and new cutoffs will be used in the real data example. Technical details and formulas can be found in Yuan and Chan (2016).

2.3. Projection Method

One major goal of MI is to test the cross-group equality of means of latent traits, especially when our interest is to study the effect of different experimental conditions or group difference. However, with the conventional approach, the test of H_τ in **Table 1** or even the estimation of $\boldsymbol{\tau}$ requires the hypothesis H_γ to hold, which is theoretically unnecessary and practically hard to achieve. In this subsection, we introduce a new setup proposed in Deng and Yuan (2016) under which the means of the latent traits can be compared even when the intercepts of manifest variables are not equal across groups. A projection method is used so that the means of manifest variables in each group are decomposed into orthogonal components of common scores and specific factors. The test of cross-group equality of the means of the common scores is essentially the test of cross-group equality of means of latent traits under the conventional setup whereas the test of cross-group equality of means of specific factors is related to but different from the test of cross-group equality of the intercepts under the conventional setup.

In the conventional setup of examining MI via Equation (1), the mean structure involves the intercepts and the means of the latent traits. The intercepts $\boldsymbol{\gamma}^{(j)}$ need to be set as equal across groups so that the means $\boldsymbol{\tau}^{(j)} = E(\boldsymbol{\xi}^{(j)})$ can be identified and estimated (Sörbom, 1974). Similarly, the means $\boldsymbol{\tau}^{(j)}$ of one group need to be set at $\mathbf{0}$ as the baseline so that the $\boldsymbol{\tau}^{(j)}$ of the other groups are the differences from those of the baseline group. To circumvent this assumption, Deng and Yuan (2016) proposed to decompose the observed variables into common scores, specific factors, and measurement errors

$$\mathbf{x}^{(j)} = \Lambda \mathbf{f}^{(j)} + \mathbf{u}^{(j)} + \mathbf{e}^{(j)},\ j = 1, \cdots, m, \tag{7}$$

where $\Lambda\mathbf{f}^{(j)}$ represents the vector of p common scores, $\mathbf{u}^{(j)}$ represents the vector of p specific factors, and $\mathbf{e}^{(j)}$ is a vector of p measurement errors, with $E[\mathbf{f}^{(j)}] = \boldsymbol{\kappa}^{(j)}$, $E[\mathbf{u}^{(j)}] = \boldsymbol{\nu}^{(j)}$, and $E[\mathbf{e}^{(j)}] = \mathbf{0}$. There is no superscript on the factor loading matrix Λ because the decomposition in Equation (7) is a step following metric invariance $H_\lambda : \Lambda^{(1)} = \cdots = \Lambda^{(m)} = \Lambda$. When metric invariance does not hold, researchers have the option to identify a subset of variables that satisfy metric invariance (Byrne et al., 1989; Millsap and Kwok, 2004). Then the projection method can be equally applied to the identified subset, as was discussed in Deng and Yuan (2016).

Note that the new setup in Equation (7) is not a simple reparameterization of the conventional setup in Equation (1). In fact, the interpretation has changed entirely. With the projection method, we assume that the space of common score is orthogonal to that of specific factors, and the comparison of means of the common scores or factors $\mathbf{f}^{(j)}$ is conducted independently from those of $\mathbf{u}^{(j)}$. Under the new setup, the mean structure of $\mathbf{x}^{(j)}$ is decomposed as

$$\boldsymbol{\mu}^{(j)} = \boldsymbol{\mu}_\kappa^{(j)} + \boldsymbol{\nu}^{(j)}, \tag{8}$$

where $\boldsymbol{\mu}_\kappa^{(j)} = \Lambda\boldsymbol{\kappa}^{(j)}$ is the part of $\boldsymbol{\mu}^{(j)} = E(\mathbf{x}^{(j)})$ that is projected onto the space of common scores, and $\boldsymbol{\nu}^{(j)}$ is the part of $\boldsymbol{\mu}^{(j)}$ that is projected onto the space of specific factors. The two components are identified once Λ is identified. Regardless of the values of $\boldsymbol{\kappa}^{(j)}$, $\boldsymbol{\mu}_\kappa^{(j)}$ is always the linear combinations of the columns of Λ.

Let $\hat{\Lambda}$ be the estimated factor loading matrix and $\bar{\mathbf{x}}^{(j)}$ be the sample means of the jth group. Then the space of the estimated common scores consists of vectors of linear combinations of the columns of $\hat{\Lambda}$, and is totally determined by $\hat{\Lambda}$. The estimated means of the common scores are consequently obtained by projecting $\bar{\mathbf{x}}^{(j)}$ onto the column space of $\hat{\Lambda}$, and we denote it as $\hat{\boldsymbol{\mu}}_\kappa^{(j)}$. Similarly, the estimated means of the specific factors are obtained by projecting $\bar{\mathbf{x}}^{(j)}$ onto the space that is orthogonal to that of $\hat{\Lambda}$, and we denote it as $\hat{\boldsymbol{\nu}}^{(j)}$. Details of the projection matrix and examples are provided in Deng and Yuan (2016). In particular, there exists $\bar{\mathbf{x}}^{(j)} = \hat{\boldsymbol{\mu}}_\kappa^{(j)} + \hat{\boldsymbol{\nu}}^{(j)}$. Also, an estimate of $\boldsymbol{\kappa}^{(j)}$ is uniquely obtained from $\hat{\boldsymbol{\mu}}_\kappa^{(j)}$, and we denote it as $\hat{\boldsymbol{\kappa}}^{(j)}$. Thus, the estimates of means of common and specific factors only depend on the sample means and estimated common factor loading matrix, and do not involve estimating the intercepts in Equation (1).

Two types of invariance tests on means can be conducted under the new setup. One test is about cross-group equality of means of common scores, which is equivalent to the test on cross-group equality of means of the latent constructs. The other test is on cross-group equality of means of specific factors. The corresponding hypotheses are

$$H_\kappa : \boldsymbol{\kappa}^{(1)} = \cdots = \boldsymbol{\kappa}^{(m)} \text{ or } \boldsymbol{\mu}_\kappa^{(1)} = \cdots = \boldsymbol{\mu}_\kappa^{(m)}, \tag{9}$$

and

$$H_\nu : \boldsymbol{\nu}^{(1)} = \cdots = \boldsymbol{\nu}^{(m)}. \tag{10}$$

The two hypotheses can also be formulated as $H_\kappa : \boldsymbol{\kappa}_d^{(j)} = \boldsymbol{\kappa}^{(j)} - \boldsymbol{\kappa}^{(1)} = \mathbf{0}$ and $H_\nu : \boldsymbol{\nu}_d^{(j)} = \boldsymbol{\nu}^{(j)} - \boldsymbol{\nu}^{(1)} = \mathbf{0}$, $j = 2, \cdots, m$. Deng and Yuan (2016) showed that $\hat{\boldsymbol{\kappa}}_d^{(j)}$ and $\hat{\boldsymbol{\nu}}_d^{(j)}$ asymptotically follow normal distributions, and each of the two hypotheses can be tested using a Wald[4] statistic $T_{gls} = NF_{gls}$ that asymptotically follows a chi-square distribution with degrees of freedom $df_\kappa = (m-1)k$ and $df_\nu = (m-1)(p-k)$, respectively. In addition to using the Wald statistics, the two hypotheses in (9) and (10) can also be tested via the bootstrap methodology, especially when the sample sizes are not large enough.

The interest of mean comparison in most studies might be to find a significant difference across groups. If this is the goal, then conventional NHT would be logically sufficient and ET is not needed. However, it is hard to imagine that the population means of different groups are literally identical. A non-significant result might be due to a small sample size and/or a small effect size. Knowing the size of the difference would be more informative even if one cares primarily about significant differences. The framework of ET would not only inform researchers the size of a possible misspecification but also provide a confidence level to it. For ET, the two hypotheses in (9) and (10) need to be reformulated, parallel to Equation (3). That is, the null hypothesis for endorsing the equality of the $\boldsymbol{\kappa}^{(j)}$, $j = 1, 2, \cdots, m$, becomes

$$H_{e\kappa} : F_{gls0} > \epsilon_0, \tag{11}$$

where F_{gls0} is the population value of F_{gls} corresponding to the T_{gls} for testing H_κ. Then the critical value for judging the significance of T_{gls} is the left-tail quantile of $\chi^2_{df_\kappa}(\delta_0)$ corresponding to level α, where $\delta_0 = N\epsilon_0$. We reject $H_{e\kappa}$ when T_{gls} is smaller than the critical value. Similarly, we can test $H_{e\nu}$ under ET, although there might be less interest in comparing the means of specific factors. As with the chi-square-difference statistics in the previous subsection, we can specify the value of ϵ_0 via $\text{RMSEA}_0 = (\epsilon_0/df_\kappa)^{1/2}$ as well as by testing $H_{e\kappa}$ using the T-size RMSEA[5] corresponding to the Wald statistic. In our package `equaltestMI`, we compute the T-size RMSEA_t corresponding to the value of the Wald statistic instead of reporting the critical value $\chi^2_{df_\kappa}(\delta_0)$. Researchers can compare the value of RMSEA_t against the adjusted cutoff values, which are printed out in the output of the R package.

A key feature of the projection method is a validity index. Let $\boldsymbol{\mu}_{d\kappa}^{(j)} = \Lambda(\boldsymbol{\kappa}^{(j)} - \boldsymbol{\kappa}^{(1)})$, $j = 2, \cdots, m$, and $\boldsymbol{\mu}_\kappa^{(d)}$ is the vector of length $p(m-1)$ formulated by stacking the $\boldsymbol{\mu}_{d\kappa}^{(j)}$; and $\boldsymbol{\nu}^{(d)}$ is the vector of length $p(m-1)$ formulated by stacking the $\boldsymbol{\nu}_d^{(j)}$, $j = 2, 3, \cdots, m$. Deng and Yuan (2016) defined a validity index for mean difference as

$$\rho_c^2 = \frac{|\boldsymbol{\mu}_\kappa^{(d)}|^2}{|\boldsymbol{\mu}_\kappa^{(d)}|^2 + |\boldsymbol{\nu}^{(d)}|^2}, \tag{12}$$

where $|\boldsymbol{\mu}_\kappa^{(d)}|^2$ and $|\boldsymbol{\nu}^{(d)}|^2$ denote the sums of squares of the elements in $\boldsymbol{\mu}_\kappa^{(d)}$ and $\boldsymbol{\nu}^{(d)}$, respectively. This validity index gives the percentage of the mean differences of the manifest variables that is due to the differences in means of the common scores. If the sample estimate $\hat{\rho}_c^2$ is not large enough, say less than 0.5, then items in the test might need to be modified or the administration of the data collection process might not be conducted properly. We will call ρ_c^2 the validity index for mean differences, because elaboration on the observed mean differences might be off the target when $\hat{\rho}_c^2$ is not sufficiently large, say greater than 0.70. In particular, when most of the mean differences in the manifest variables are not due to those in the latent traits, the validity of the measurements might be questionable. Then the empirical meaning of the observed differences will be different from the truth, which will create interpretational confounding. The extent to which the observed mean differences reflect the mean differences of the latent variables is not available following the analysis of the mean structures in the conventional setup, where cross-group equality of intercepts is a prerequisite for estimating mean differences of latent variables.

3. REAL DATA EXAMPLE

In this section, we introduce the R package `equaltestMI` and illustrate its use via a real data example. Both ET and the projection method are implemented in the R package, which is available on CRAN and can be used on any R platform with version 3.1.0 or above. The development of `equaltestMI` relies on R packages `lavaan` (Rosseel, 2012) for obtaining chi-square statistics of invariance tests and `semTools` (semTools Contributors, 2016) for computing chi-square-difference tests and fit indices. The function for computing adjusted RMSEA cutoff values for ET is adapted from the R codes available at http://www3.nd.edu/~kyuan/mgroup/Equivalence-testing.R. The input to `equaltestMI` can be either raw data sets with group membership indicator or sample means and covariances.

3.1. Data Set

Literacy-related difficulties for many children are due to lack of exposure to print or instructional resources, and thus socioeconomic status (SES) is an important demographic variable that strongly relates to academic achievement. The data we use for the illustration are from Lee and Al Otaiba (2015), and their Table 1 contains sample statistics (sample

[4] Wald statistics are typically formulated via generalized least squares (GLS), and are commonly called GLS statistics in the psychometric literature.

[5] Note that the ncp δ_0, RMSEA_0 and RMSEA_t corresponding to the Wald statistic T_{gls} might be different from that corresponding to T_{ml} for a given condition of misspecification. But their difference is tiny unless the model is severely misspecified.

sizes, means, covariances) on early literacy skills from 2 sociodemographic groups of kindergartners, with $N_1 = 78$ boys ineligible for free or reduced-price lunch (FRL) and $N_2 = 174$ boys eligible for FRL. The interest of Lee and Al Otaiba is whether measurements on literature proficiency are invariant when compared students with lower SES (eligible for FRL) against those with higher SES (ineligible for FRL). There are six manifest variables in measuring literacy constructs: (1) letter-name fluency, (2) letter-sound fluency, (3) blending, (4) elision, (5) real words spelling, and (6) pseudo-words spelling. Following from Snow's (2006) definition of componential skills and the work of Schatschneider et al. (2004) on National Early Literacy Panel (NELP), the six variables aim to measure three aspects of literacy constructs: (1) alphabet knowledge, which refers to children's familiarity with letter forms, names, and corresponding sounds; (2) phonological awareness, which encompasses the ability to detect, manipulate, or analyze sounds in spoken language in varying complexities such as words, syllables, and phonemes; and (3) spelling, which measures the ability to spell words with letters (Piasta and Wagner, 2010). As indicated in **Figure 1**, alphabet knowledge, phonological awareness, and spelling are the three latent constructs behind the six variables. Lee and Al Otaiba (2015) examined the MI issues using the conventional methods. Let the boys who are ineligible for FRL be group one and those eligible for FRL be group two. We will use the six-variable-two-group model to illustrate the application of the new methods via the package `equaltestMI`.

3.2. Package `equaltestMI`

To use `equaltestMI` for the first time, one needs to download the package from CRAN and load it into R environment. This can be done by entering the following commands in R:

```
install.packages(equaltestMI)                                   1
library(equaltestMI)                                            2
```

The line numbers in the right margin are for convenience of explaining the codes in our illustration, not part of the R commands. After loading the package `equaltestMI` into R, there is no need to load `lavaan` and `semTools` separately since they are listed as dependent packages of `equaltestMI`. However, one does need to have the two packages installed before the library command on line 2, otherwise `equaltestMI` cannot be successfully loaded.

The package `equaltestMI` has multiple R functions. The one that is routinely used is `eqMI.main()`. Other functions can be used to test the cross-group equality of population covariance matrices or to obtain adjusted RMSEA cutoff values in a separate analysis. Interested users are referred to supplementary material (http://www3.nd.edu/~kyuan/eqMI/Supplementary_Material_MI.pdf) and the page of `equaltestMI` on CRAN (https://cran.r-project.org/web/packages/equaltestMI/

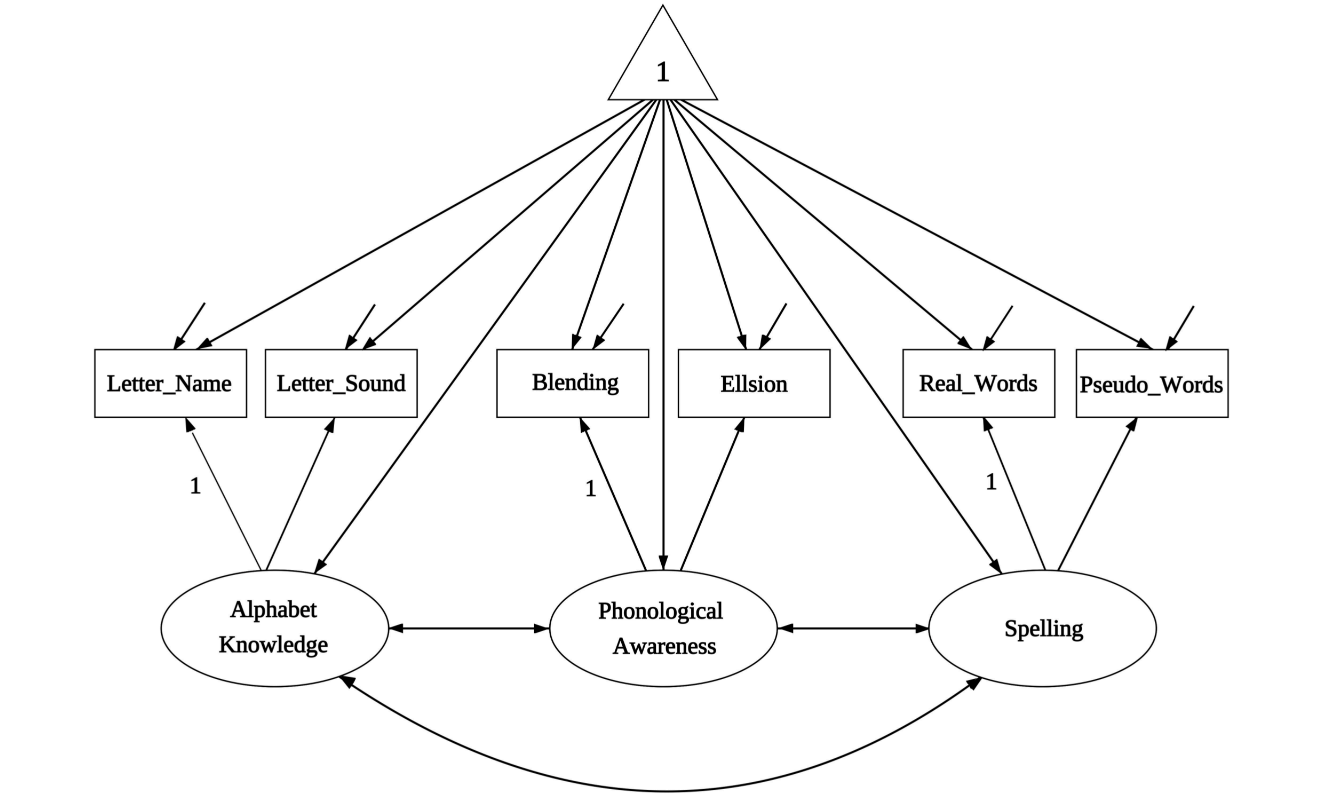


FIGURE 1 | The path diagram for the model of Lee and Al Otaiba (2015).

index.html), where details for using different functions are documented.

Different arguments can be provided to `eqMI.main()` for customized analysis. However, data input (raw data or sample means and covariances) has to be in required format. For a raw data set, column represents variables and row represents observations, and an additional column of duplicated numbers for group membership is needed for all the involved samples. For either raw data or sample statistics, the first row needs to be variable names, and including a group-membership indicator. If input data are sample statistics, the sample means must be stored in the format of vectors immediately following the variables names; and sample covariance matrix stored in the format of matrix are next, see **Appendix A** for the format. In particular, the first row of each file must have a space before the first variable name. The label `’mean’` must be included prior to the numerical values of the mean vector. In addition, variable names are also needed in the first column of the covariance matrix as required by `lavaan`, these will be further used to in the model syntax, to be presented below.

For conducting the tests of MI, we first need to import the sample means and sample covariance matrices into R. This is done by the following R codes for this example:

```
setwd("C:/research/equaltestMI")                      3
Group1 <- read.table(’Group1.txt’, header = TRUE)     4
Group2 <- read.table(’Group2.txt’, header = TRUE)     5
Group1 <- as.matrix(Group1)                           6
Group2 <- as.matrix(Group2)                           7
M1 <- Group1[1,]                                      8
M2 <- Group2[1,]                                      9
Cov1 <- Group1[2:7,]                                  10
Cov2 <- Group2[2:7,]                                  11
```

The code `setwd("C:/research/equaltestMI")` on Line 3 sets the working directory as the folder where the data files are stored. The `read.table` command on Lines 4 and 5 put the sample means and sample covariance matrices of the two groups in `Group1.txt` and `Group2.txt` into R environment. The format of the two data files `Group1.txt` and `Group2.txt` are provided in **Appendix A**. The argument `header` is set to be `TRUE` in order to identify the variable names that are needed to set the model. Lines 6 and 7 then use `as.matrix` to convert the formats of `Group1` and `Group2` to `matrix` as required by `lavaan`, so that the sample means and covariance matrices extracted from `Group1` and `Group2` are in the correct formats. Lines 8 to 11 separate the sample means from the sample covariance matrices for each group according to the positions of the values in the data files.

Another argument that is needed by `eqMI.main()` is the model statement. Since `lavaan` and `semTools` are used to compute chi-square and chi-square-difference test statistics, the model syntax is written following the convention of `lavaan`:

```
model <- ’                                            12
AlphabetKnowledge =~ Letter_Name+ Letter_Sound        13
PhonologicalAwareness =~ Blending + Elision           14
Spelling =~ Real_Words + Pseudo_Words                 15
’                                                     16
```

where the single quotation marks (can also be double quotation marks) enclose a model statement. The sign `=~` is used to indicate the relationship between a latent factor and its manifest indicators/variables. On the left of each `=~` is the label for a latent factor and those following `=~` are the corresponding manifest variables that loaded onto the latent factor. The manifest variables that load onto the same latent factor are connected by "+." The names of the latent factors in the model statement cannot duplicate any of the variable names in `Group1.txt` or `Group2.txt`.

To perform ET and the projection method for MI with two groups, we supply the following arguments to `eqMI.main()`:

```
test <- eqMI.main(model = model,                      17
        sample.nobs = c(78, 174),                     18
        sample.mean = list(M1, M2),                   19
        sample.cov = list(Cov1, Cov2),                20
        meanstructure = TRUE,                         21
        output = ’both’,                              22
        quiet = FALSE,                                23
        equivalence.test = TRUE, adjRMSEA = TRUE,     24
        projection = TRUE, bootstrap = FALSE)         25
```

The `model` on Line 17 is the SEM model we defined using the convention of `lavaan` on Lines 12 to 16. The `sample.nobs` on Line 18 contains the numbers of observations for the two groups, and more numbers are needed with more groups. The `sample.mean` on Line 19 is a list of sample means obtained on Lines 8 and 9, and `sample.cov` is a list of sample covariance matrices. The `meanstructure = TRUE` on Line 21 is needed if mean structures are involved instead of saturated means. The `output = ’both’` on Line 22 requires the results of tests of both the mean and covariance structures (steps 1 to 6c in **Table 1**) be printed out. One can also output the results of only the mean structure or the covariance structure by specifying `output = ’mean’` or `output = ’covariance’`. The `quiet = FALSE` on Line 23 tells the program to print out a summary to R console that contains test statistics and fit measures of all the involved tests as described in **Table 1**. The arguments `equivalence.test = TRUE` and `adjRMSEA = TRUE` on Line 24 tell the program to conduct ET and print out T-size RMSEA and adjusted cutoff values. The arguments `projection = TRUE` and `bootstrap = FALSE` on Line 25 tell the program to conduct mean comparison using the projection method. Bootstrap resampling is not invoked in this example due to the absence of raw data. However, bootstrap can be enabled to obtain empirical *p*-values for the tests of equalities of common and specific factors using the projection method once raw data become available, and the details are documented in the online supplementary material.

3.3. Output

```
---------- Equality of Population Covariance Matrices under NHT ----------
               Chisq Df       pvalue
fit.pop.cov 48.85008 21 0.0005261139

---------- Chi-Square and Chi-Square-Difference Test under NHT ----------
                        Chisq Df   pvalue Chisq.diff Df.diff   pvalue
fit.pop.cov            48.850 21    0.001
fit.configural.g1       4.408  6    0.622
fit.configural.g2      10.641  6    0.100
fit.combine.groups     15.049 12
fit.metric             20.033 15    0.171      4.984       3    0.173
fit.residuals          42.512 21    0.004     22.479       6    0.001
fit.varfactor          54.175 27    0.001     11.663       6    0.070
fit.scalar             23.732 18    0.164      3.699       3    0.296
fit.strong.means       41.066 21    0.006     17.334       3    0.001
fit.strict.residuals   45.968 24    0.004     22.237       6    0.001
fit.strict.means       63.630 27    0.000     17.662       3    0.001

-------------- T-size epsilon, RMSEA, and Adjusted Cutoff Values under ET --------------
                    epsilon_t  RMESA_t   cut.01   cut.05   cut.08   cut.10 goodness-of-fit
fit.pop.cov            0.209    0.141    0.076    0.097    0.121    0.139            poor
fit.configural.g1      0.028    0.097    0.116    0.133    0.157    0.175       excellent
fit.configural.g2      0.071    0.154    0.116    0.133    0.157    0.175            fair
fit.metric             0.049    0.181    0.151    0.164    0.187    0.205            fair
fit.residuals          0.140    0.216    0.116    0.133    0.157    0.175            poor
fit.varfactor          0.078    0.161    0.116    0.133    0.157    0.175        mediocre
fit.scalar             0.040    0.163    0.151    0.164    0.187    0.205           close
fit.strong.means       0.125    0.289    0.151    0.164    0.187    0.205            poor
fit.strict.residuals   0.138    0.215    0.116    0.133    0.157    0.175            poor
fit.strict.means       0.127    0.291    0.151    0.164    0.187    0.205            poor

------ Means of Latent and Specific Factors by the Projection Method and under NHT ------
                 Chisq Df       pvalue
fit.mvmean   22.388932  6 0.0010292280
fit.common   19.433779  3 0.0002223618
fit.specific  4.015387  3 0.2598074102
Validity Index is 0.9885648

------  Means of Latent and Specific Factors by the Projection Method and under ET ------
             epsilon_t  RMESA_t   cut.01   cut.05   cut.08   cut.10 goodness-of-fit
fit.mvmean       0.139    0.215    0.116    0.133    0.157    0.175            poor
fit.common       0.137    0.302    0.151    0.164    0.187    0.205            poor
fit.specific     0.042    0.168    0.151    0.164    0.187    0.205            fair

---------- Cross-group Comparison of Latent Factor Means ----------
                       latent_1 latent_2 latent_d    SE_d      z_d
AlphabetKnowledge     39.20010 34.77505 -4.42505 1.87963 -2.35422
PhonologicalAwareness 10.50104  8.29014 -2.21090 0.59194 -3.73503
Spelling              22.14624 17.69643 -4.44981 1.11260 -3.99946

---------- Cross-group Comparison of Common Scores ----------
             common_1 common_2 common_d    SE_d      z_d
Letter_Name  39.20010 34.77505 -4.42505 1.87963 -2.35422
Letter_Sound 45.65332 40.49980 -5.15351 2.18906 -2.35422
Blending     10.50104  8.29014 -2.21090 0.59194 -3.73503
Elision       7.11369  5.61597 -1.49772 0.40099 -3.73503
Real_Words   22.14624 17.69643 -4.44981 1.11260 -3.99946
Pseudo_Words 16.45361 13.14762 -3.30600 0.82661 -3.99946

---------- Cross-group Comparison of Specific Factors ----------
           specific_1 specific_2 specific_d    SE_d      z_d
Letter_Name   6.05990    6.54495    0.48505 0.92562  0.52403
Letter_Sound -5.20332   -5.61980   -0.41649 0.79478 -0.52403
Blending      0.40896    0.78986    0.38090 0.21495  1.77204
Elision      -0.60369   -1.16597   -0.56228 0.31730 -1.77204
Real_Words    1.73376    1.54357   -0.19019 0.25533 -0.74490
Pseudo_Words -2.33361   -2.07762    0.25600 0.34367  0.74490
```

Running the R codes on Lines 17 to 25 generates the above output that has eight parts. Part 1 (Lines 26 to 28) contains the results of testing equality of population covariance matrices under NHT. The package `lavaan` does not provide such a test so that we developed an R function `eqMI.covtest()` to perform this test using the method of Lagrange multiplier. Part 2 (Lines 30 to 42) contains the results of MI under the conventional NHT, including the chi-square and chi-square-difference test statistics along with their degrees of freedom and p-values. Part 3 of the output (Lines 44 to 55) are the results of MI under ET, consisting of the T-size ϵ_t, RMSEA$_t$, adjusted cutoff values, and labels of the goodness of fit by comparing RMSEA$_t$ against the adjusted cutoff values. Note that the results on Line 49 to 55 are based on the chi-square-difference statistics whereas those on Lines 46 to 48 are based on the T_{ml} statistic as reported in Parts 1 and 2.

Part 4 of the output (Lines 57 to 62) contains the results of testing the cross-group equality of means using the projection method and under NHT. The numbers following `fit.mvmean` is the results of the Wald test of equality of means of the manifest variables, those following `fit.common` and `fit.specific` are the results of the Wald tests of the cross-group equality of means of common and specific factors, respectively. Line 62 contains the value of the validity index according to Equation (12). Part 5 (Lines 64 to 68) contains the results of mean comparison by the projection method and under ET, where the T-size ϵ_t and RMSEA$_t$ are based on the Wald statistics reported in Part 4. For each of the tests listed in the output, one can extract details such as parameter estimates and standard errors from the resulting R object `test` on Line 17.

Parts 6 to 8 (Lines 70 to 92) of the output of `eqMI.main()` contain parameter estimates, standard errors and the corresponding z-scores under the projection approach. Those corresponding to the differences of the estimates across groups are also included. These are the same under NHT and ET.

4. RESULTS

It follows from Line 28 that the equality of population covariance matrices is rejected under NHT at level $\alpha = 0.05$. According to the results on Line 46, we cannot regard the two population covariance matrices as equal under ET unless we can tolerate a model with RMSEA$_t$ $=$ 0.141. With the adjusted cutoff value for poor model being at 0.139, the model under equal covariance matrices is worse than poor. Consequently, we reject the hypothesis and conclude that the two population covariance matrices cannot be regarded as equal.

We next turn to the components of the measurement models as represented by **Figure 1**. Under conventional NHT, Lines 33 and 34 indicate that the significance level of the statistic T_{ml} for group 1 (boys ineligible for FRL) is 0.622, and for group 2 (boys eligible for FRL) is 0.100. One would conclude that configural invariance holds in the population under NHT and move to the next step of the analysis. In contrast, under ET, the goodness of fit for group 1 (Line 47) is excellent with RMSEA$_t = 0.097$; but that for group 2 (Line 48) is fair with RMSEA$_t$ $=$ 0.154. Configural invariance is again established under the condition that we are able to tolerate a model with fair fit or RMSEA$_t = 0.154$.

Moving to the next analysis of metric invariance (cross-group equality of factor loading matrices H_λ) under NHT (Line 36), the p-value corresponding to the chi-square difference statistic of 4.984 is 0.173, and we conclude that metric invariance holds and move to the next step of the analysis of MI. Under ET, the results on Line 49 indicated that RMSEA$_t$ $=$ 0.181 and the goodness of fit is fair. Metric invariance is endorsed only we can accept a model of misspecification with RMSEA$_t = 0.181$ beyond that in configural invariance.

Following metric invariance, we can next test cross-group equality of variance components (error variances and factor variances-covariances; steps 4a and 5a in **Table 1**). Alternatively, we can also move to test scalar invariance and cross-group equality of means of latent constructs (steps 4b–6c in **Table 1**).

Under conventional NHT, with a p-value of 0.001 on Line 37, the chi-square-difference statistic suggests that the hypothesis H_ψ is unlikely to hold. Under ET, results on Line 50 indicate that error variances may not be regarded as equal across the two groups unless we can tolerate a poor model with T-size RMSEA$_t = 0.216$.

Move to the mean structure under NHT (Line 39), with a p-value of 0.296 for the chi-square-difference statistic, one would conclude that scalar invariance holds in the population. Under ET (Line 52), the T-size RMSEA corresponding to the chi-square-difference statistic for scalar invariance is 0.163, and the model achieved close fit when compared RMSEA$_t$ against the adjusted cutoff values.

Under NHT, results on Lines 40 to 42 imply that we cannot endorse the cross-group equality of means of the latent constructs (H_τ) nor that of error variances (H_ψ). Thus, strong invariance is achieved but not strict invariance. Results under ET (Lines 53 to 55) also suggest that strict invariance does not hold unless we can tolerate poor models with RMSEA$_t$ being above 0.20.

Results on Line 59 is the Wald test for cross-group equality of means of the 6 manifest variables. The results for testing the cross-group equality of means of the common and specific factors by the projection method under NHT (Lines 60 and 61) indicate that the two groups have different means of common factors but their means in specific factors might be equal. Consequently, 98.9% of the squared mean differences for manifest variables is due to mean differences in the three latent constructs: alphabet knowledge, phonological awareness, and spelling, indicating that the six variables are good measures of the literacy skills. The results following the projection method under ET (Lines 67 and 68) indicate that we can endorse H_{ev} and regard the means of the specific factors as being equal across the two groups if a misspecification with RMSEA$_t$ $=$ 0.168 is tolerable, or be able to accept a fair model. However, we will have to accept a poor model in order to endorse $H_{e\kappa}$ or to tolerate a misspecification with RMSEA$_t = 0.302$.

Lines 70 to 92 of the output are the results for the means of the latent, common and specific factors, following the projection approach. Those on Lines 70 to 74 indicate that boys eligible for FRL have significantly smaller means of latent traits. As expected, the two groups are significantly different in the mean of each of the six common scores, with those in the low-SES group being uniformly smaller. In contrast, the two SES groups do not show significant differences on any of the six specific factors, implying that most of the cross-group differences in manifest variables are due to those in latent traits.

For this example, the conventional method of NHT endorses both metric invariance and scalar invariance. However, NHT cannot claim that the two properties hold in the population, since it is designed for rejecting the null hypothesis instead of proving that the null hypothesis holds. In contrast, the method of ET did not conclude cross-group equality of either the factor loadings or intercepts. Instead, ET claims that, with probability of 0.95, the difference between the two factor-loading matrices is less than 0.049 as measured by F_{ml} or less than 0.181 as measured by RMSEA. Similarly, ET claims that, with probability of 0.95, the difference between the two vectors of intercepts is less than 0.040 as measured by F_{ml} or less than 0.163 as measured by RMSEA. With the projection method, ET claims that with probability 0.95 the two vectors of means of specific factors differ by less than 0.042 as measured by F_{gls} or less than 0.168 as measured by the corresponding RMSEA. We are able to endorse metric and scalar invariance only if we can tolerate models with fair fit, and the endorsement is attached with a T-size and a probability.

While the statistic T_{ml} is not significant for the hypothesis of scalar invariance H_{γ} in the example, it is rare in practice. The projection method allows us to estimate and compare the means of latent traits as long as metric invariance is endorsed, and a validity index is also provided.

5. CONCLUSION

In this article, we introduced two recently proposed methods, combined the projection-based method and ET, implemented the new methods in an R package, and illustrated the use of the R package via a real data example. We believe that the development will contribute to the use of the cutting-edge methodology in substantive areas where MI is needed in group comparison. In particular, we recommend that researchers report the results of ET together with those under NHT even if they may not want to abandon the method of NHT in studying MI.

We only illustrated ET in the context of MI in this article. ET is equally applicable in other contexts where NHT has been the dominant methodology, especially in areas where models are needed to account for the relationship among the observed variables (e.g., growth curve modeling, time series analysis, item response models) rather than rejecting the null hypotheses. Recent developments for ET in structural equation modeling include Marcoulides and Yuan (2017) and Yuan et al. (2016), where both RMSEA and CFI (Bentler, 1990) can be used for determining the tolerable size of misspecification. ET can also be used for parameter testing, especially when a particular value of the parameter is of special interest (Wellek, 2010).

Throughout the article, we have used RMSEA to quantify the cross-group difference in model parameters. However, Cohen's *d* or standardized mean difference is regularly used in *t*-test and ANOVA. We might adopt Cohen's *d* for ET when quantifying the cross-group differences in the means of latent traits. However, it is not clear how to generalize the standardized mean difference to multiple groups when the covariance matrices of the latent traits are heterogeneous. Correlated latent factors might also cause difficulty with interpretation if we generalize *d* to a multivariate version (Huberty, 2002). Vandenberg and Lance (2000) discussed the pros and cons of different approaches to mean comparison and recommend using overall model fit indices to assess the appropriateness of imposed invariance constraints.

Like any statistical methodology, ET needs a statistic that approximately follows a central/non-central chi-square or another distribution of known form. When such a distribution is not available, especially when conditions are not met (e.g., non-normally distributed data, missing values), alternative statistics other than T_{ml} might be needed. Bootstrap methodology can also be considered. Further developments are needed in these directions.

AUTHOR CONTRIBUTIONS

GJ carried out the project, write the example and the initial draft of the article. YM did the program and coding of the software `equaltestMI`. KY directed the project, and finalized the article in writing.

ACKNOWLEDGMENTS

The research was supported by the National Science Foundation under Grant No. SES-1461355.

REFERENCES

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychol. Bull.* 107, 238–246. doi: 10.1037/0033-2909.107.2.238

Byrne, B. M. (2010). *Structural Equation Modeling with AMOS: Basic Concepts, Applications, and Programming*. New York, NY: Routledge.

Byrne, B. M., Shavelson, R. J., and Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychol. Bull.* 105, 456–466. doi: 10.1037/0033-2909.105.3.456

Deng, L., and Yuan, K.-H. (2016). Comparing latent means without mean structure models: a projection-based approach. *Psychometrika* 81, 802–829. doi: 10.1007/s11336-015-9491-8

Gorsuch, R. L. (1983). *Factor Analysis*, 2nd Edn. Hillsdale, NJ: Lawrence Erlbaum.

Harman, H. H. (1976). *Modern Factor Analysis,* 3rd Edn. Chicago, IL: University of Chicago Press.

Huberty, C. J. (2002). A history of effect size indices. *Educ. Psychol. Measur.* 62, 227–240. doi: 10.1177/0013164402062002002

Horn, J. L., and McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Exp. Aging Res.* 18, 117–144. doi: 10.1080/03610739208253916

Horn, J. L., McArdle, J. J., and Mason, R. (1983). When is invariance not invarient: a practical scientist's look at the ethereal concept of factor invariance. *South. Psychol.* 1, 179–188.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika* 36, 409–426. doi: 10.1007/BF02291366

Kim, E. S., Kwok, O.-M., and Yoon, M. (2012). Testing factorial invariance in multilevel data: a monte carlo study. *Struct. Equ.*

Model. Multidiscipl. J. 19, 250–267. doi: 10.1080/10705511.2012.659623

Lai, K., and Green, S. B. (2016). The problem with having two watches: assessment of fit when RMSEA and CFI disagree. *Multivar. Behav. Res.* 51, 220–239. doi: 10.1080/00273171.2015.1134306

Lee, J. A. C., and Al Otaiba, S. (2015). Socioeconomic and gender group differences in early literacy skills: a multiple-group confirmatory factor analysis approach. *Educ. Res. Eval.* 21, 40–59. doi: 10.1080/13803611.2015.1010545

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

Marcoulides, K. M., and Yuan, K.-H. (2017). New ways to evaluate goodness of fit: a note on using equivalence testing to assess structural equation models. *Struct. Equ. Model. Multidiscipl. J.* 24, 148–153. doi: 10.1080/10705511.2016

Marsh, H. W., Guo, J., Parker, P. D., Nagengast, B., Asparouhov, T., Muthén, B., et al. (2017). What to do when scalar invariance fails: the extended alignment method for multi-group factor analysis comparison of latent means across many groups. *Psychol. Methods* 21, 405–426. doi: 10.1037/met0000113

Mellenbergh, G. J. (1989). Item bias and item response theory. *Int. J. Educ. Res.* 13, 127–143. doi: 10.1016/0883-0355(89)90002-5

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika* 58, 525–543. doi: 10.1007/BF02294825

Millsap, R. E., and Kwok, O.-M. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychol. Methods* 9, 93–115. doi: 10.1037/1082-989X.9.1.93

Millsap, R. E. (2011). *Statistical Approaches to Measurement Invariance*. New York, NY: Routledge.

Piasta, S. B., and Wagner, R. K. (2010). Developing early literacy skills: a meta-analysis of alphabet learning and instruction. *Read. Res. Q.* 45, 8–38. doi: 10.1598/RRQ.45.1.2

Rosseel, Y. (2012). lavaan: an R package for structural equation modeling. *J. Statis. Softw.* 48, 1–36. doi: 10.18637/jss.v048.i02

Schatschneider, C., Fletcher, J. M., Francis, D. J., Carlson, C., and Foorman, B. R. (2004). Kindergarten prediction of reading skills: a longitudinal comparative analysis. *J. Educ. Psychol.* 96, 265–282. doi: 10.1037/0022-0663.96.2.265

semTools Contributors (2016). *semTools: Useful Tools for Structural Equation Modeling*. R package version 0.4-14. Available online at: https://CRAN.R-project.org/package=semTools

Snow, C. E. (2006). "What counts as early literacy in early childhood?" in *Blackwell Handbook of Early Childhood Development*, eds K. McCartney and D. Phillips (Malden, MA: Blackwell), 274–294.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structures between groups. *Br. J. Math. Statist. Psychol.* 27, 229–239. doi: 10.1111/j.2044-8317.1974.tb00543.x

Steiger, J. H. (1998). A note on multiple sample extensions of the RMSEA fit index. *Struct. Equ. Model. Multidiscipl. J.* 5, 411–419. doi: 10.1080/10705519809540115

Steiger, J. H., and Lind, J. C. (1980). "Statistically-based tests for the number of common factors," in *Paper Presented at the Annual Meeting of the Psychometric Society* (Iowa City, IA).

Vandenberg, R. J., and Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. *Organ. Res. Methods* 3, 4–70. doi: 10.1177/109442810031002

Wellek, S. (2010). *Testing Statistical Hypotheses of Equivalence and Noninferiority,*(2nd Edn.,) Boca Raton, FL: Chapman & Hall/CRC Press LLC.

Yuan, K.-H., and Bentler, P. M. (2004). On chi-square-difference and z tests in mean and covariance structure analysis when the base model is misspecified. *Educ. Psychol. Measure.* 64, 737–757. doi: 10.1177/0013164404264853

Yuan, K.-H., and Bentler, P. M. (2006). Mean comparison: Manifest variable versus latent variable. *Psychometrika* 71, 139–159. doi: 10.1007/s11336-004-1181-x

Yuan, K.-H., and Chan, W. (2016). Measurement invariance via multigroup SEM: Issues and solutions with chi-square-difference tests. *Psychol. Methods* 21, 405–426. doi: 10.1037/met0000080

Yuan, K.-H., Chan, W., Marcoulides, G. A., and Bentler, P. M. (2016). Assessing structural equation models by equivalence testing with adjusted fit Indexes. *Struct. Equat. Model.* 23, 319–330. doi: 10.1080/10705511.2015.1065414

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

APPENDIX A

Data files `Group1.txt` and `Group2.txt` used in the real data example

```
Group1.txt:
             Letter_Name Letter_Sound Blending  Elision Real_Words Pseudo_Words
Mean            45.26000     40.45000 10.91000  6.51000   23.88000   14.12000
Letter_Name    207.36000    159.09696 32.58864 25.80480   61.77600   45.07488
Letter_Sound   159.09696    280.22760 42.88788 36.74765   76.12348   60.20374
Blending        32.58864     42.88788 18.23290 10.71258   19.05103   14.21910
Elision         25.80480     36.74765 10.71258 20.07040   20.37235   16.70861
Real_Words      61.77600     76.12348 19.05103 20.37235   73.61640   47.42852
Pseudo_Words    45.07488     60.20374 14.21910 16.70861   47.42852   44.35560
```

```
Group2.txt:
             Letter_Name Letter_Sound Blending  Elision Real_Words Pseudo_Words
Mean            41.32000      34.8800  9.080000  4.450000  19.24000   11.07000
Letter_Name    295.84000     232.2000 38.995840 20.173880  67.59256   57.77136
Letter_Sound   232.20000     324.0000 43.164000 22.824000  77.95440   60.45840
Blending        38.99584      43.1640 19.009600  9.260204  23.42802   16.27152
Elision         20.17388      22.8240  9.260204 10.048900  15.25404   11.04174
Real_Words      67.59256      77.9544 23.428024 15.254040  64.32040   38.41099
Pseudo_Words    57.77136      60.4584 16.271520 11.041744  38.41099   38.68840
```

frontiers in Psychology

METHODS
published: 15 May 2018
doi: 10.3389/fpsyg.2018.00740

The Impact of Partial Measurement Invariance on Testing Moderation for Single and Multi-Level Data

Yu-Yu Hsiao[1]* and *Mark H. C. Lai*[2]

[1] *Center on Alcoholism, Substance Abuse, and Addictions, University of New Mexico, Albuquerque, NM, United States,*
[2] *School of Education, University of Cincinnati, Cincinnati, OH, United States*

Moderation effect is a commonly used concept in the field of social and behavioral science. Several studies regarding the implication of moderation effects have been done; however, little is known about how partial measurement invariance influences the properties of tests for moderation effects when categorical moderators were used. Additionally, whether the impact is the same across single and multilevel data is still unknown. Hence, the purpose of the present study is twofold: (a) To investigate the performance of the moderation test in single-level studies when measurement invariance does not hold; (b) To examine whether unique features of multilevel data, such as intraclass correlation (ICC) and number of clusters, influence the effect of measurement non-invariance on the performance of tests for moderation. Simulation results indicated that falsely assuming measurement invariance lead to biased estimates, inflated Type I error rates, and more gain or more loss in power (depends on simulation conditions) for the test of moderation effects. Such patterns were more salient as sample size and the number of non-invariant items increase for both single- and multi-level data. With multilevel data, the cluster size seemed to have a larger impact than the number of clusters when falsely assuming measurement invariance in the moderation estimation. ICC was trivially related to the moderation estimates. Overall, when testing moderation effects with categorical moderators, employing a model that accounts for the measurement (non)invariance structure of the predictor and/or the outcome is recommended.

Keywords: measurement equivalence, measurement invariance, moderation, interaction effects, structural equation modeling, hierarchical linear modeling, multilevel modeling

Edited by:
Ehri Ryu,
Boston College, United States

Reviewed by:
Jason C. Immekus,
University of Louisville, United States
Dexin Shi,
University of South Carolina,
United States

****Correspondence:***
Yu-Yu Hsiao
yuyuhsiao.office@gmail.com

Specialty section:
This article was submitted to Quantitative Psychology and Measurement,
a section of the journal Frontiers in Psychology

Received: *31 July 2017*
Accepted: *27 April 2018*
Published: *15 May 2018*

Citation:
Hsiao Y-Y and Lai MHC (2018) The Impact of Partial Measurement Invariance on Testing Moderation for Single and Multi-Level Data.
Front. Psychol. 9:740.
doi: 10.3389/fpsyg.2018.00740

Many theories in education and psychology rely on moderators, which in Baron and Kenny's (1986) words, "[affect] the direction and/or strength of the relation between an independent or predictor variable and a dependent or outcome variable" (p. 1,174). For many years, social and behavioral researchers are interested in understanding whether a specific moderation effect occurs as well as what factors may influence the extent of the moderation effect. Numerous methodological studies regarding different aspects of moderation effects have been done in contexts such as multiple regression (Aiken and West, 1991), multiple-group structural equation modeling (multiple-group SEM; Jaccard and Wan, 1996), latent variable models with observed composites (Bohrnstedt and Marwell, 1978; Busemeyer and Jones, 1983; Hsiao et al., 2018), within-subject designs (Judd et al., 1996, 2001), cross-level interactions (Kreft et al., 1995), and Bayesian estimations (Lüdtke et al., 2013).

Much of the methodological research regarding moderation effects focused on continuous variables, and less research has been done for categorical moderators. As an example of the latter, researchers may be interested in how the effect of social support on happiness differs by gender. Gender as a categorical variable is treated as the moderator, and social support and happiness are the predictor and outcome variables, respectively. In testing such a moderation with conventional methods such as multiple regression and multiple-group SEM, researchers implicitly assume that the predictor and the outcome variables are measurement invariant across the categorical moderators; that is, the measurement characteristics for social support and happiness are the same by different gender categories. However, such an assumption is seldom investigated before testing moderation effects. Additionally, little is known about how measurement non-invariance influences the estimation of the moderation effects. Hence, it is worth investigating whether measurement invariance for both the predictor and the outcome variables with respect to the moderator categories is a necessary prerequisite before conducting a moderation effect testing.

Measurement invariance (MI) is an important issue in a variety of social and behavioral research settings, especially when the data are collected from multiple populations (Millsap and Kwok, 2004). Full MI holds when individuals with identical ability but from different groups have the same propensity to get a particular score on that specific ability scale (Yoon and Millsap, 2007). Under the multiple-group confirmatory factor analysis framework, a simplified but commonly used version of MI analyses can be conducted by testing four models with hierarchical orders across groups: equal model structures (configural invariance), equal factor loadings (metric invariance), equal intercepts (scalar invariance), and equal unique factor variances (strict invariance; Vandenberg and Lance, 2000; Millsap and Kwok, 2004; Chen et al., 2005; Brown, 2015). Among the four types of MI, metric invariance has been suggested as one basic requirement for doing prediction (Vandenberg and Lance, 2000), which is closely related to moderation effect as moderation effect is about the difference in path coefficients across groups. Hence, in this paper we focus on the impact of metric non-invariance on the estimation of moderation effects. We also focus on testing moderation effects with the multiple-group approach, which is generally being used for examining measurement invariance.

PREVIOUS RESEARCH ON THE EFFECT OF METRIC NON-INVARIANCE ON PREDICTION

Millsap (1995, 1997, 1998, 2007) delineated several theorems and corollaries for the relationship between MI and prediction bias. Donahue (2006) conducted a simulation study to examine the change of the prediction accuracy when the measure of the exogenous (predictor) variable was non-invariant in some part of the factor loadings, or with the presence of partial metric invariance, across groups. Her study found that, if one correctly assumes a partial invariance model on the latent predictors' structures, the path coefficient estimates on the outcome variables are unbiased even with a larger degree of metric non-invariance (i.e., more non-invariant items) on the latent predictors. However, the study only included the effects on tests of simple regression coefficient in each group, but not moderation, which can be defined as the difference in path coefficients across groups. Additionally, the study did not show the consequences of failing to correctly model the non-invariance structure.

Guenole and Brown (2014) used Monte Carlo studies to investigate the impact of ignoring measurement invariance (including metric invariance) on testing linear and nonlinear effects (including moderation effects). They adopted relative bias of the estimated path coefficients and 95% coverage rate of the estimated confidence intervals from both the reference group and focal group. They found biased estimates of the path coefficients from the two groups when two or more (out of six) ignored non-invariant loadings occurred. The same results were observed when the non-invariance occurred for predictors and outcomes[1].

In the present research, we address two gaps from the work of Donahue (2006) and Guenole and Brown (2014). First, we would show the degree to which estimations and tests of moderation are affected when researchers incorrectly assume that (metric) MI holds. Second, we are interested in whether the location of measurement non-invariance, particularly in the predictor or in the outcome variable, makes a difference. Furthermore, we extend their work by investigating the Type I error rate of misidentifying null moderation effect and the statistical power of detecting nonzero moderation effects in the presence of non-invariance.

Additionally, Donahue (2006) and Guenole and Brown (2014) focused on single level data structure, in which all the observations were assumed to be independent. However, educational and psychological data often have nesting structures (e.g., students nested within classrooms; Kim et al., 2012). For example, a researcher is interested in how the association between students' motivation and their academic achievement differs in public and private schools. Since students are nested within schools, the school variable is a moderator defined in the between level and motivation is a predictor defined in the within level. Therefore, the scenario represents a "cross-level" moderation effects. In this situation, the measurement characteristics of motivation and academic achievement are assumed invariant across school types (i.e., public vs. private). It is still unclear that how multilevel measurement metric (non)invariance across groups in the between level influences the cross-level moderation effects. Therefore, we also show how unique features of multilevel data affect the MI-moderation relationship[2].

[1] One prerequisite to interpret moderation effects in the presence of non-invariance is that the constructs being measured are still conceptually comparable across groups (i.e., configural invariance). If the predictor and/or outcome represent different measurement structure across groups, the computed moderation effect may not be meaningful.

[2] Throughout this article, we investigated the multilevel measurement (non)invariance across an explicitly defined grouping variable, not the non-invariance for the between-versus within-level. Additionally, we assumed that the outcome variable is defined at the within-level but not at the between-level. Finally, the grouping variable was defined at the between-level but not the within-level in

STUDY 1

In Study 1, we aim to show the effect of measurement non-invariance on the power and Type I error rate when testing a moderator with two categories. Both the predictor and the outcome have a measurement structure and the moderation effects are tested with multiple-group approach, as shown in **Figure 1**. Specifically,

$$\begin{aligned}\mathbf{X}_g &= \boldsymbol{\lambda}_{Xg}F_{Xg} + \boldsymbol{\delta}_g,\\ \mathbf{Y}_g &= \boldsymbol{\lambda}_{Yg}F_{Yg} + \boldsymbol{\varepsilon}_g,\\ F_{Yg} &= \gamma_g F_{Xg} + \zeta_g,\end{aligned}$$

where $g = 1, 2$ was the group index number, $\mathbf{X} = [X_1, X_2, \ldots]'$ and $\mathbf{Y} = [Y_1, Y_2, \ldots]'$ were observed indicators as shown in **Figure 1**, $\boldsymbol{\lambda}_X$ and $\boldsymbol{\lambda}_Y$ were two vectors of factor loadings of the indicators on the latent variables, $\boldsymbol{\delta}$ and $\boldsymbol{\varepsilon}$ were vectors of the effects of unique factor on $\mathbf{X}$ and $\mathbf{Y}$, γ_g is the path coefficient between F_X and F_Y for group g, and ζ was the latent disturbance term for F_Y. In addition, both the impacts of having metric non-invariance on the outcome and on the predictor were investigated. The simulation study was described below.

Monte Carlo Simulation

The study had a 3 (p_{ni}, number of non-metric-invariant indicators) × 4 ($\boldsymbol{\gamma} = \{\gamma_1, \gamma_2\}'$, vector of population regression coefficients of the two groups) × 2 (location of non-invariance) × 2 (N, sample size of each group) design. In each condition there were two groups, and the sample sizes were assumed equal across groups. Both the predictor F_X and the outcome F_Y were latent variables with six indicators.

Number of Non-metric-Invariant Indicators, p_{ni}

Across the simulation conditions, p_{ni} will either be 0, 2, or 4. For all indicators in Group 1, the factor loadings were set to 0.7, while some of those in Group 2 were set to 0.3 to represent moderate degree of metric non-invariance. This was similar to the conditions in some previous studies (Kaplan and George, 1995; Donahue, 2006).

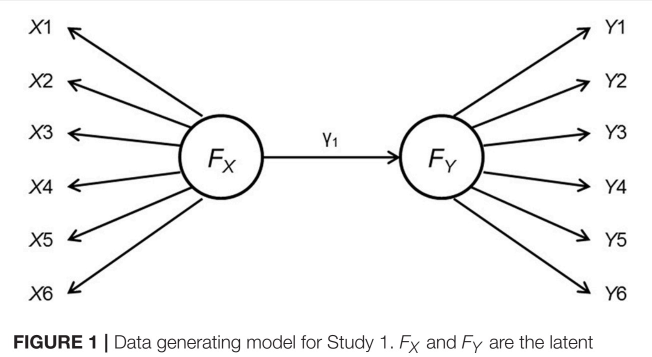


FIGURE 1 | Data generating model for Study 1. F_X and F_Y are the latent predictor and outcome variables, each indicated by six observed indicators.

Regression Coefficients, γ

There were four levels of $\boldsymbol{\gamma}$, two of which with equal regression coefficients ($\{0.1, 0.1\}$ and $\{0.5, 0.5\}$) and two with them different ($\{0.5, 0.33\}$ and $\{0.33, 0.5\}$). In the equal $\boldsymbol{\gamma}$ conditions the grouping variable did not moderate the effects of F_X on F_Y, and Type I error rates were investigated. We were also interested in whether the effect of F_X being large (i.e., 0.5) and small (i.e., 0.1) influences Type I error rates. In the conditions with different $\boldsymbol{\gamma}$ the effects of F_X on F_Y were different for Group 1 and for Group 2, so there were moderation effects between groups and F_X on F_Y and powers of detecting the true moderation effects were investigated. The numbers were chosen based on the benchmark of small ($\gamma = 0.1$), medium ($\gamma = 0.33$), and large effects ($\gamma = 0.5$; Cohen, 1988).

Location of Non-invariance

The metric non-invariance occurred either on only F_X or only F_Y. Note that this design factor were not applicable to conditions with $p_{ni} = 0$.

Sample Size, N

There were two levels of sample size: 200 and 500, in consistent with some previous studies (e.g., Yoon and Millsap, 2007).

Mplus 7.0 (Muthén and Muthén, 2012) was used to generate 500 data sets for each condition. All variables were assumed multivariate normally distributed. The two factor variances in Group 1 were 1.0 and those in Group 2 were 1.3. For Group 1, the unique factor variances of all indicators were set to 0.51 in the population, so that the invariant indicators had a variance of 1.0. The unique factor variances for Group 2 were set to $0.51 \times 1.3 = 0.663$ so that the proportion of explained variances for the invariant indicators was constant across groups. Because scalar invariance was not the focus of the present study and might not be required for correctly modeling moderation effects, all intercepts and factor means in the population were set to zero.

The data sets generated were then analyzed in Mplus. The analytic model was identified by fixing the factor loadings of the first indicators for F_X and for F_Y to the population value (i.e., 0.7), while allowing the latent factor variances of F_X and of F_Y to be freely estimated. Hence, both F_X and F_Y were scaled to the same unit as the population model and across replications, so that the $\boldsymbol{\gamma}$ values from the two groups were comparable. To identify the mean structure, the latent factor mean of F_X and the latent intercept of F_Y were fixed to zero for both groups, while the intercepts and the unique factor variances were allowed to be freely estimated without cross-group equality constraints, as scalar and strict invariance conditions were not assumed.

For conditions with $p_{ni} = 0$, the data sets were analyzed by fitting only the model with metric invariance. For other conditions with $p_{ni} > 0$, both the (misspecified) model with metric invariance and the (correct) model with partial metric invariance were fitted. Then for each data set, we obtained the point estimate of $\Delta\hat{\gamma} = \gamma_1 - \gamma_2$ (using the MODEL

the multilevel case. All the mentioned conditions not investigated in the present study may have different implications that are also worthy of future studies.

`CONSTRAINT` command in Mplus) and the Wald test statistic (using the `MODEL TEST` command in Mplus) for the null hypothesis $\gamma_1 = \gamma_2$. Note that we also obtained the results for the likelihood ratio test, which is usually more accurate for finite samples, but we only presented the results for the Wald test as the two tests were nevertheless asymptotically equivalent and produced similar empirical powers and Type I error rates across simulation conditions.

The dependent variables of investigation for the simulations were the percentage of replications where the test statistics were statistically significant at 0.05 level and the standardized bias of $\Delta\hat{\gamma}$. If in the population, $\gamma_1 = \gamma_2$, then the percentage of replications with statistically significant Wald test statistic was the empirical Type I error rate (α^*). Taking into account the sampling variability in 500 replications, an α^* between 3.4% and 7.3% is within the 95% confidence interval when the true Type I error rate is 5%. Empirical Type I error rates over the range of [3.4%, 7.3%] are defined as biased. We expected to see biased Type I error rates and the standardized biases to be large when metric invariance is incorrectly assumed.

If in the population $\gamma_1 \neq \gamma_2$, the percentage where the test statistics were statistically significant at 0.05 level was the empirical power. Given that power is a function of effect size and sample size, the empirical power rates yielded from fitting the model with metric invariance in $p_{\text{ni}} = 0$ condition were treated as the baseline; those yielded with $p_{\text{ni}} > 0$ from incorrectly assuming measurement invariance and correctly assuming partial invariance models were then compared to the baseline. We expected to see power estimates from models incorrectly assuming measurement invariance were more different from the baseline then the correctly assuming partial invariance models.

Denote $\hat{\gamma}_1^{(i)}$ and $\hat{\gamma}_2^{(i)}$ as the estimated values of γ_1 and γ_2 for the ith replication, and $\bar{\gamma}_1$ and $\bar{\gamma}_2$ as the corresponding means across replications. The standardized bias (Collins et al., 2001) was computed as

$$\text{standardized bias} = \frac{(\bar{\gamma}_1 - \bar{\gamma}_2) - (\gamma_1 - \gamma_2)}{SD(\hat{\gamma}_1 - \hat{\gamma}_2)},$$

where

$$SD(\hat{\gamma}_1 - \hat{\gamma}_2) = \sqrt{\frac{\sum_{i=1}^{R}[(\hat{\gamma}_1^{(i)} - \hat{\gamma}_2^{(i)}) - (\bar{\gamma}_1 - \bar{\gamma}_2)]^2}{R}},$$

and $i = 1, 2, \ldots, R$ was the index of replications where $R = 500$. The standardized bias was the ratio of the average raw bias over the standard error of the sample estimator of the parameter, and a standardized bias with absolute value < 0.40 was regarded as acceptable (Collins et al., 2001).

Result

The simulation results for the condition with null moderation effects were displayed in **Table 1**. When the measurement invariance assumption held on both the predictor and the outcome in population model (i.e., $p_{\text{ni}} = 0$), using the analytic model assuming measurement invariance across groups yielded unbiased moderation effect estimates and unbiased α^*.

When the non-invariance occurred on F_X, as partial metric invariance was the correctly specified model, with a partial invariance model α^* was close to the 0.05 nominal significance level and the moderation effect was estimated with absolute values of standardized bias <0.02 ($<$ 0.40 as acceptable). On the other hand, α^* was inflated when metric invariance were falsely assumed. The difference between α^* from the nominal level increased as one or more of p_{ni}, N, and the values of $\boldsymbol{\gamma}$ increased. For example, when $N = 200$, $p_{\text{ni}} = 2$, and $\boldsymbol{\gamma} = \{0.1, 0.1\}$, $\alpha^* = 4.2\%$; when $N = 500$, $p_{\text{ni}} = 4$, and $\boldsymbol{\gamma} = \{0.1, 0.1\}$,

TABLE 1 | Empirical type I error rate (in percentage) and standardized bias for study 1.

			Non-invariance on F_X				**Non-invariance on F_Y**			
			Type I error (%)		**Std. Bias ($\Delta\hat{\gamma}$)**		**Type I error (%)**		**Std. Bias ($\Delta\hat{\gamma}$)**	
N	**γ**	**p_{ni}**	**MI**	**pMI**	**MI**	**pMI**	**MI**	**pMI**	**MI**	**pMI**
200	{0.1, 0.1}	0	4.2	–	0.00	–	–	–	–	–
		2	4.2	4.6	−0.12	0.00	4.2	4.0	0.11	0.01
		4	6.0	4.6	−0.34	−0.01	5.4	4.2	0.33	0.02
	{0.5, 0.5}	0	4.6	–	0.01	–	–	–	–	–
		2	8.0	4.4	−0.64	0.00	7.8	4.2	0.62	0.02
		4	35.2	4.0	−1.60	−0.01	38.8	4.8	1.74	0.02
500	{0.1, 0.1}	0	4.6	–	0.01	–	–	–	–	–
		2	5.4	5.2	−0.18	0.01	5.0	5.2	0.19	0.01
		4	7.6	5.0	−0.51	0.01	7.0	5.0	0.52	0.00
	{0.5, 0.5}	0	4.0	–	−0.01	–	–	—	–	–
		2	14.4	3.6	−1.06	−0.01	13.8	4.8	0.97	−0.01
		4	77.8	3.2	−2.79	−0.01	77.8	5.2	2.74	−0.01

p_{ni}, number of non-metric-invariant indicators; $\boldsymbol{\gamma}$, population regression coefficient of F_Y on F_X; MI, analytic model assumed metric invariance; pMI, analytic model correctly assumed partial metric invariance; Std. Bias, standardized bias = $\Delta\hat{\gamma}/SD(\Delta\hat{\gamma})$, where $SD(\Delta\hat{\gamma})$ is the standard deviation of the differences in the estimated γs across all replications.

$\alpha^* = 7.6\%$; and when $N = 500$, $p_{\text{ni}} = 4$, and $\boldsymbol{\gamma} = \{0.5, 0.5\}$, $\alpha^* = 77.8\%$. An analysis of variance (ANOVA) including N, $\boldsymbol{\gamma}$, and p_{ni} showed that p_{ni} produced the largest impact on α^* ($\eta^2 = 0.34$), followed by $\boldsymbol{\gamma}$ ($\eta^2 = 0.21$) and N ($\eta^2 = 0.04$). The bias of the estimated values of $\Delta\gamma$ followed a similar pattern. For instance, With $N = 500$, $p_{\text{ni}} = 4$, and $\boldsymbol{\gamma} = \{0.5, 0.5\}$, the standardized bias of the null moderation effects was −2.79, which was a substantial bias.

The pattern of α^* and the absolute values of the standardized bias when non-invariance occurred in F_{Y} was very similar to those when non-invariance occurred in F_{X}. However, the sign of the standardized bias was reversed, which means that when non-invariance occurred in the outcome's structure, the moderation effects were overestimated. Considering both the locations of the non-invariance, we found that using models that incorrectly assumed measurement invariance would result in substantially biased moderation effect estimate and inflated Type I error rate.

Table 2 showed the results of both the powers and standardized biases with nonzero moderation effects. When the non-invariance occurred on F_{X}, the corrected partial metric invariance models performed well as they showed no bias on the moderation effect estimates with standardized biases from −0.03 to 0.01. On the contrary, the metric invariance model yielded biased estimates of the moderation effects and the influence was more salient as both N and p_{ni} increased. For example, when $\boldsymbol{\gamma} = \{0.5, 0.33\}$, the standardized bias was −0.43 with $N = 200$ and $p_{\text{ni}} = 2$; the standardized bias increased to −1.84 with $N = 500$ and $p_{\text{ni}} = 4$. An ANOVA showed that p_{ni} produced the largest impact on the biased moderation estimates ($\eta^2 = 0.79$), followed by N ($\eta^2 = 0.09$) and $\boldsymbol{\gamma}$ ($\eta^2 = 0.01$).

In terms of the powers for detecting the moderation effects, the corrected partial invariance model yielded powers around 30% and 60% for N equals 200 and 500, respectively. Such power estimates were close to population model with the measurement invariance assumption held (33% for $N = 200$ and 70% for $N = 500$). On the other hand, if metric invariance was falsely assumed, there was a substantial decrease in powers for the conditions where non-invariance occurred. For example with $\boldsymbol{\gamma} = \{0.5, 0.33\}$, $N = 500$, $p_{\text{ni}} = 2$, and non-invariance on F_{X}, the empirical power was half as would be obtained when metric invariance held in the population (33.8% vs. 70.2%); with $\boldsymbol{\gamma} = \{0.33, 0.5\}$, $N = 200$, $p_{\text{ni}} = 4$, and non-invariance on F_{Y}, the empirical power was only 1/8 as the power would be obtained when metric invariance held in the population (4.2% vs. 33.0%).

Note that power loss was detected as both the N and p_{ni} increased when the non-invariance occurred on F_{X} and $\boldsymbol{\gamma} = \{0.5, 0.33\}$; simulation conditions related to non-invariance occur-ed on F_{Y} and $\boldsymbol{\gamma} = \{0.33, 0.5\}$ would lead to inflated power estimates as both the N and p_{ni} increased. The main reason for different patterns on the power estimates were that when the factor loadings of Group 1 (0.7) was larger than those of Group 2 (0.3) in the presence of non-invariance on F_{X}, the estimated moderation effect was negatively biased, whereas when non-invariance occurred in F_{Y}, the estimated moderation effect was positively biased. Additionally, the true moderation effect was −0.17 when $\boldsymbol{\gamma} = \{0.33, 0.5\}$; therefore, the negative biases caused by falsely assuming measurement invariance would result in more negative moderation effects estimates and inflated power.

STUDY 2

In Study 2, we aim to extend the scope of the MI-moderation relation to multilevel data. We focused on how the measurement (non-)invariance across groups at the between level influences the test of cross-level moderation effect, which was one of the prevailing issues among social and behavioral research.

TABLE 2 | Empirical power (in percentage) and standardized bias for study 1.

			Non-invariance on F_{X}				**Non-invariance on F_{Y}**			
			Power (%)		**Std. Bias ($\Delta\hat{\gamma}$)**		**Power (%)**		**Std. Bias ($\Delta\hat{\gamma}$)**	
N	**γ**	**p_{ni}**	**MI**	**pMI**	**MI**	**pMI**	**MI**	**pMI**	**MI**	**pMI**
200	{0.5, 0.33}	0	32.2	–	−0.01	–	–	–	–	–
		2	16.0	30.8	−0.43	−0.02	53.0	31.2	0.48	0.00
		4	4.4	31.2	−1.12	−0.03	81.8	29.4	1.29	0.02
	{0.33, 0.5}	0	33.0	–	0.02	–	33.0	–	–	–
		2	50.0	30.8	−0.61	0.01	15.8	30.0	0.52	0.03
		4	77.0	26.8	−1.57	−0.01	4.2	27.4	1.52	0.04
500	{0.5, 0.33}	0	70.2	–	−0.01	–	–	–	–	–
		2	33.8	67.2	−0.69	−0.02	89.0	67.4	0.76	−0.01
		4	5.0	62.4	−1.84	−0.02	99.2	62.4	2.01	−0.02
	{0.33, 0.5}	0	67.6	–	0.00	–	–	–	–	–
		2	90.6	67.2	−1.01	−0.01	36.0	66.2	0.79	0.00
		4	99.4	59.4	−2.71	−0.01	5.0	62.8	2.38	0.00

p_{ni}, number of non-metric-invariant indicators; $\boldsymbol{\gamma}$, population regression coefficient of F_{Y} on F_{X}; MI, analytic model assumed metric invariance; pMI, analytic model correctly assumed partial metric invariance; Std. Bias, standardized bias = $\Delta\hat{\gamma}/SD(\Delta\hat{\gamma})$, where $SD(\Delta\hat{\gamma})$ is the standard deviation of the differences in the estimated γs across all replications.

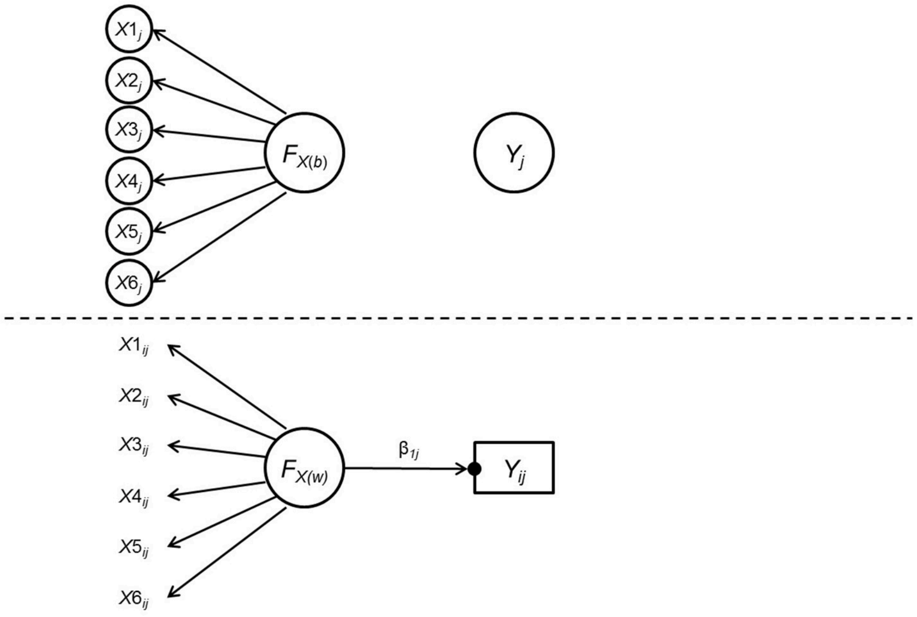


FIGURE 2 | Data generating model for Study 2. $F_{X(w)}$ and $F_{X(b)}$ are the latent predictor variable at the within-level and the between-level, respectively. Y_{ij} and Y_j are the within-level and the between-level components of the outcome variable Y. β_{1j} = within-level regression coefficient of Y on $F_{X(w)}$, whose magnitude varies across clusters as indicated by the black dot. Conditioning on the grouping variable, measurement invariance was assumed across clusters such that the within-level and the between-level factors loadings were identical, and that there were no residual variances for the six indicators at the between-level.

Specifically, we used the data generating model shown in **Figure 2**, which was one of the simplest models including multilevel measurement (non-)invariance and a within-level predictor, to depict the cross-level moderation effect. As can be seen in **Figure 2**, the latent predictor was measured by six indicators and the cross-level interaction effect was denoted by the difference between the within-level path coefficient from the predictor to the outcome across groups. It was assumed that the predictor did not have an effect on the outcome in the between level.

Because multilevel data are usually of larger sample size, we expect the impact of multilevel non-invariance on the Type I error rate and power to be bigger. In addition, we are interested in whether the impact varies across multilevel specific design factors such as the intraclass correlation (ICC), number of clusters, and cluster size. Because in Study 1, we found that different locations of non-invariance mainly resulted in changes in signs of the biases of the moderation effects, in Study 2 we only focused on measurement non-invariance on the predictor side. Likewise, we only consider the positive moderation effects condition in Study 2 given that negative moderation effect led to similar results in biases in Study 1. A second Monte Carlo simulation study was conducted, as described below.

Monte Carlo Simulation

The study had a 2 (p_{ni}) × 2 ($\boldsymbol{\gamma}$) × 2 (ICC, intraclass correlation) × 2 (m, number of clusters) × 2 (c, cluster size) design. In each condition there were two groups (Group 1 and Group 2), and sample sizes (both within and between) were assumed equal across groups. The latent predictor, F_{X}, had the same six-indicator measurement structure in both the within and the between level as in Study 1; the observed outcome, Y_{ij}, of the ith observation in the jth cluster, contained no measurement error and was assumed measurement invariant. The model can be expressed as

$$\begin{aligned}\text{Within level: } Y_{ijg} &= Y_{(b)jg} + \gamma_{10g}F_{\mathrm{X}(w)ijg} + r_{ijg},\\ \mathbf{X}_{ijg} &= \mathbf{X}_{(b)jg} + \boldsymbol{\lambda}_{(w)g}F_{\mathrm{X}(w)ijg} + \boldsymbol{\delta}_{(w)ijg};\\ \text{Between level: } Y_{(b)jg} &= \gamma_{00g} + \zeta_{jg},\\ \mathbf{X}_{(b)jg} &= \alpha_g + \boldsymbol{\lambda}_{(b)g}F_{\mathrm{X}(b)jg} + \boldsymbol{\delta}_{(b)jg},\end{aligned}$$

where $\mathbf{X} = [X_1, X_2, \ldots]'$ was a vector containing the observed values of the indicators, and their group means comprised $\mathbf{X}_{(b)}$. The vectors $\boldsymbol{\lambda}_{(w)}$ and $\boldsymbol{\lambda}_{(b)}$ contained the within-level and between-level factor loadings, respectively. In this study we assumed that $\boldsymbol{\lambda}_{(w)g} = \boldsymbol{\lambda}_{(b)g} = \boldsymbol{\lambda}_g$. In the within level, $F_{\mathrm{X}(w)}$ (the within-level exogenous factor) had an effect of magnitude γ_{10g} on Y, where g is the group index. In the between level, $F_{\mathrm{X}(b)}$ (the between level exogenous factor) had no effect on Y. Note that there were no between-level random effects on γ_{10g} and on the factor loadings. We also assumed measurement invariance across clusters, implying that $\boldsymbol{\delta}_{(b)jg} = \mathbf{0}$ and homogeneous $\boldsymbol{\delta}_{(w)jg}$ across clusters (Jak et al., 2013), in addition to $\boldsymbol{\lambda}_{(w)g} = \boldsymbol{\lambda}_{(b)g}$. The design factors were described below.

Number of Non-metric-Invariant Indicators, p_{ni}

p_{ni} was either 0 or 2 out of the six indicators of F_{X}. Whereas two of the factor loadings were always set to 0.7 in Group 1, for conditions with $p_{\text{ni}} = 2$, those loadings were set to 0.3 in Group 2. The factor loadings for other four indicators were 0.7, 0.3, 0.5, and 0.6, for both the within level and the between level.

Regression Coefficients, γ

There were two levels of $\boldsymbol{\gamma}$: $\{0.3, 0.2\}$ (moderation present) and $\{0.3, 0.3\}$ (moderation absent).

Intraclass Correlation, ICC

Based on previous simulations (Kim et al., 2012), in this study there were two levels of ICC: 0.10 and 0.35, representing small and large within-cluster correlations for the latent variable F_{X} and for the outcome Y.

Cluster Size, *c*

Based on previous literature (Clarke, 2008; Kim et al., 2012), there were two levels of cluster size: 5 and 20, representing small and medium number of observations within a cluster. For simplicity we generated data with all clusters having the same size in both groups.

Number of Clusters, *m*

Hox and Maas (2001) suggested the number of groups larger than 100 as the minimum requirement for yielding accurate multilevel regression estimates. Later on, Maas and Hox (2005) found groups number equal to 30 could also yield accurate multilevel regression estimates. McNeish (2017) did a literature review on 70 multilevel studies and found 90% of them fail to meet Hox and Maas's criterion of 100 clusters, and that the median number of clusters was 44. In the present study, we specified the number of clusters in each group either 30 or 100, representing the small and large number of clusters.

Mplus 7.0 was used to generate and analyze (with `ESTIMATOR=MLR`) 500 data sets for each condition. All exogenous variables and random effects were assumed multivariate-normally distributed. For both groups the variances of $F_{\text{X}(w)}$ and Y both equaled to 1.0, and that of $F_{\text{X}(b)}$ and ζ_{jg} were functions of the ICC. The variance of $\boldsymbol{\delta}_{(w)ijg}$ was set to $0.5\mathbf{I}$ so that the level-1 unique factor variances were similar in values to those in Study 1 (i.e., 0.51 in Study 1 when the latent factor variance is one). The covariance and mean structure were identified similarly as in Study 1 by fixing the factor loadings of the first indicators for $F_{\text{X}(b)}$ to the population value and the latent mean of $F_{\text{X}(b)}$ to zero for both groups. Additionally, within the same group the factor loadings were constrained to be equal in the between and the within levels so that metric invariance was assumed across clusters (Jak et al., 2013). Because scalar invariance was not the focus of the present study and may not be required for correctly modeling moderation effects, all intercepts and factor means in the population were set to zero.

The dependent variables of investigation were the standardized biases and the rejection rates of the Wald test statistics for the difference in γ_{10}, which reflected either the empirical Type I error rate (α^*) or the empirical power, and were obtained in the same manner as in Study 1. Also as in Study 1, for conditions with $p_{\text{ni}} = 2$, both the metric invariance model and the partial metric invariance model were fitted. We expected that model falsely assuming measurement invariance would lead to biased moderation estimation, inflated Type I error rate (when $p_{\text{ni}} = 0$), and power more different from the baseline (when $p_{\text{ni}} = 2$).

Result

Results for Study 2 were shown in **Table 3**. In the conditions absent of moderation effects ($\boldsymbol{\gamma} = \{0.3, 0.3\}$), fitting data with a metric invariance model when the true population model followed the measurement invariance assumption ($p_{\text{ni}} = 0$) led to unbiased moderation effect estimates and unbiased α^*, regardless of the level of ICC, *m*, and *c*. The same pattern was observed while employing the corrected partial metric invariance model to fit data from a measurement non-invariance population, as such practice also led to unbiased moderation estimates and α^* close to the 5% nominal significance level across different ICC, *m*, and *c* simulation conditions.

When non-invariance occurred ($p_{\text{ni}} = 2$), fitting data with a metric invariance model yielded substantially underestimated moderation effect and inflated α^*. Such trend became more salient as *m* and *c* increased. For example with ICC = 0.10, $p_{\text{ni}} = 2$, $m = 30$, and $c = 5$, the standardized bias was -0.55 with α^* of 9%; with ICC = 0.10, $p_{\text{ni}} = 2$, $m = 100$, and $c = 20$, the standardized bias increased to -2.13 with α^* of 55.2%. An ANOVA analysis on the standardized bias with p_{ni}, *m*, *c*, and ICC showed that p_{ni} had the largest impact on estimation biases ($\eta^2 = 0.70$), followed by *c* ($\eta^2 = 0.08$), *m* ($\eta^2 = 0.06$), and ICC (η^2 close to 0). ICC showed no impact of falsely assuming metric invariance on yielding biased moderation estimates and inflated α^*. For example with ICC = 0.10, $p_{\text{ni}} = 2$, $m = 100$, and $c = 5$, the standardized bias was -1.07 with α^* of 16.4%; increasing the ICC to 0.35 while keeping the other design factors to be the same led to similar results with standardized bias = -1.04 and $\alpha^* = 15.8\%$.

In the conditions with nonzero moderation effect ($\boldsymbol{\gamma}$: $\{0.3, 0.2\}$), again, employing the correctly specified partial invariance model resulted in unbiased moderation effect estimates. On the other hand, falsely assuming metric invariance led to substantially underestimated moderation effects across simulation conditions with standardized biases from -0.39 to -1.61. Consistent with the null moderation condition, an ANOVA analysis on the standardized bias indicated p_{ni} had the largest impact on estimation biases ($\eta^2 = 0.70$), followed by *c* ($\eta^2 = 0.10$), *m* ($\eta^2 = 0.06$), and ICC (η^2 close to 0). There was also a substantial loss in power when fitting a metric invariance model to data draw from a population with non-invariance. For example, the power of the simulation condition of ICC = 0.10, $p_{\text{ni}} = 0$, $m = 100$, and $c = 20$ was 80% but it dropped to 16.8% when ICC = 0.10, $p_{\text{ni}} = 2$, $m = 100$, and $c = 20$. Again, ICC only had a trivial effect on the deflation of power.

DISCUSSION

In the literature, the impact of measurement invariance on testing moderation effects has not been fully examined. The ratio of the non-invariant items have been found to be an important factor on the estimation accuracy of the path coefficients by the moderating groups (e.g., Guenole and Brown, 2014). In much of

TABLE 3 | Empirical type I error rate, power, and standardized bias for study 2.

				$\gamma = \{0.3, 0.3\}$				$\gamma = \{0.3, 0.2\}$			
				Type I Error (%)		**Std. Bias ($\Delta\hat{\gamma}$)**		**Power (%)**		**Std. Bias ($\Delta\hat{\gamma}$)**	
ICC	p_{ni}	m	c	**MI**	**pMI**	**MI**	**pMI**	**MI**	**pMI**	**MI**	**pMI**
0.10	0	30	5	6.60	–	0.01	–	15.00	–	0.01	–
			20	7.20	–	−0.02	–	34.80	–	−0.02	–
		100	5	5.80	–	0.02	–	31.20	–	0.02	–
			20	6.00	–	0.04	–	81.00	–	0.03	–
	2	30	5	9.00	5.80	−0.55	0.01	7.00	14.20	−0.41	0.01
			20	25.60	7.00	−1.17	−0.02	10.00	34.40	−0.91	0.02
		100	5	16.40	6.40	−1.07	0.01	4.80	28.60	−0.81	0.02
			20	55.20	5.80	−2.13	0.03	16.80	79.20	−1.60	0.03
0.35	0	30	5	6.20	–	0.02	–	15.20	–	0.02	–
			20	7.40	–	−0.02	–	35.20	–	−0.03	–
		100	5	5.40	–	0.01	–	30.80	–	0.02	–
			20	5.40	–	0.03	–	80.60	–	−0.03	–
	2	30	5	7.60	5.20	−0.53	0.02	6.80	14.60	−0.39	0.02
			20	25.20	7.20	−1.16	−0.02	10.20	32.80	−0.91	−0.02
		100	5	15.80	5.80	−1.04	0.01	4.60	25.80	−0.79	0.02
			20	54.60	5.40	−2.13	0.03	18.20	78.20	−1.61	0.03

γ, population regression coefficient of F_Y on F_X; ICC, intraclass correlation; p_{ni}, number of non-metric-invariant indicators; m, number of clusters; c, number of observations in a cluster; MI, analytic model assumed metric invariance; pMI, analytic model assumed partial metric invariance; Std. Bias, standardized bias = $\Delta\hat{\gamma}/SD(\Delta\hat{\gamma})$, where $SD(\Delta\hat{\gamma})$ is the standard deviation of the differences in the estimated γs across all replications.

previous work, the focus was limited to single level data structure, without considerations of nested data structure. Additionally, the direct statistical test of the moderation effect was largely ignored in previous research. The current study investigated the impact of partial measurement invariance, with a focus on the metric invariance, on the estimation and testing of moderation effects on both single and multilevel structures, in terms of standardized bias, power and Type I error rate.

The results suggest that incorrectly assuming metric invariance holds while estimating moderation effects would lead to biased estimates. The impact is more salient as the number of non-invariant items increases, which is consistent with Guenole and Brown (2014)'s and Shi et al. (2017)'s findings with direct effects. On the other hand, fitting models correctly assuming partial metric invariance yielded accurate estimates regardless of samples size, main effects, number of non-invariant items, and the location of the non-invariance occurred.

In testing null moderation effects (i.e., γs are equal between two groups), the high Type I error rate yielded from falsely assuming metric invariance is not only related to the non-invariant item ratio but the magnitude of the main effects. These results suggest that evaluation of measurement invariance is of more importance when the main effect of the predictor is larger. On the other hand, the Type I error rates were on or below 5% with models correctly assuming partial metric invariance. Thus, before examining moderation effects, the metric invariance assumption should not be presumed without conducting any invariance test, even for cases in which the moderation tests turn out to be non-significant.

The location of the non-invariance (predictor vs. outcome) is associated with the direction of the biases of the moderation effects. In our simulations, all of the non-invariance conditions were specified such that factor loadings of Group 1 were equal or larger than those of Group 2. As evident from the simulation results, under such settings ignoring predictor non-invariance leads to underestimation of the moderation effects, whereas ignoring outcome non-invariance results in overestimated moderation effects. Our findings are consistent to Chen (2008) and Guenole and Brown (2014), in which they found that non-invariance on the predictor with lower factor loadings in group 2 would lead to underestimated path coefficient in group 1 (γ_1) and overestimated path coefficient in group 2 (γ_2). Hence, the moderation effect ($\gamma_1 - \gamma_2$) would likely be underestimated. On the other hand, non-invariance on the outcome changes the association to the opposite direction and results in the overestimation of the moderation effects.

Compared with models correctly assuming partial metric invariance, models falsely assuming metric invariance yielded moderation test with statistical power varying substantially. Taking into account the signs of the moderation effects, when the moderation effects are positive, ignoring non-invariance on the predictors leads to power loss, but ignoring non-invariance on the outcomes leads to increased power (at the cost of highly inflated Type I error rate). Likewise, an opposite association between the location of non-invariance and power is observed when the moderation effects are negative. Therefore, the increase in power in half of our simulation conditions in **Table 2** is actually a byproduct of sacrificing the estimation accuracy of the moderation effects (i.e., overestimation). Ignoring non-invariance and resulting in power gain or loss depends on (a) the location of the non-invariance, (b) the signs of the moderation effects. Overall, it is not recommended to fit a model assuming metric invariance when the assumption is actually violated, even though it may increase the power of the moderation test.

There is also prospective evidence that falsely assuming multilevel metric invariance across groups has a negative impact on the estimation of the cross-level moderation effects, which leads to either substantially inflated Type I error rate or inflated/deflated statistical power of the moderation test. Both increases in m and in c, or in other words an increase in the total sample size, resulted in bigger problems in the estimation accuracy as well as α^* and power. Thus, for multilevel data, even with only one-third of the indicators being non-metric-invariant, tests of moderation can become hugely misleading.

Across simulation conditions, the number of non-invariant items played a huge role in influencing the performance of the moderation estimates. Researchers use multilevel data with a number of clusters (e.g., number of classrooms) larger than 100 or cluster size larger than 20 (e.g., 20 students in each classroom) should be particularly cautious about the negative impact of non-invariant items. Additionally, the cluster size (c) seemed to have a larger impact than the number of clusters (m) when falsely assuming measurement invariance in the moderation estimation. Intraclass correlation (ICC) was trivially related to the moderation estimates, probably because the path of interest was defined in the within-level. On the contrary, previous research has shown that ICC is highly related to between-level analysis (Kim et al., 2012). Thus, one potential explanation for the discrepancy is that, given that the cross-level moderation coefficients were mainly defined in the within-level, the level of data dependency has less influence on the moderation effect estimates.

Findings from Study 1 and Study 2 highlight the importance of testing metric invariance before conducting a moderation test with both single and multilevel data structure. If the metric invariance assumption is violated, a partial metric invariance model in which the non-invariant factor loadings between groups are correctly reflected should be employed. Researchers should also be aware that the MI-moderation relationship is highly affected by the ratio of non-invariant items in the scale and the overall sample size. Overall, while testing moderation effects in a multiple-group analysis setting, we recommend the test of measurement invariance for both the predictors and outcomes by the moderator groups. If the measurement invariance assumption holds, then employing models with such an assumption implied is appropriate. On the other hand, if the measurement invariance assumption is violated, then the use of a corrected partial invariance model would yield more accurate estimates and unbiased Type I error rates or power.

Some limitations and future study directions should be addressed. First, the research scenario only focused on metric invariance (i.e., invariance of the factor loadings). In practice, non-invariance may exist in the intercepts, factor loadings, unique factor variances, or some combinations of them. Previous simulation studies on latent growth modeling have shown that ignoring intercept non-invariance only leads to biased factor mean (or intercept) estimates (Kim and Willson, 2014). Research on multiple-group analysis also showed that ignoring intercept non-invariance has less impact on the prediction bias of the path coefficient in each group (Guenole and Brown, 2014). Therefore, we suspect that the impact of intercept non-invariance on the moderation effect estimates should be much smaller than that of factor loading non-invariance, but more conclusive evidence needs to be obtained from future methodological inquiries.

Second, for Study 2 we only tested cross-level moderation in the present study, but moderation effects may also occur at the between level, in which factors such as ICC may play a more important role in affecting the moderation estimates. Lastly, in the simulation, the indicators were assumed to be continuous and normal distributed when conditioned on the latent factors. It is important to see how measurement non-invariance with skewed and categorical indicators influence the estimation of the moderation effects. Therefore, future study can investigate the impact of falsely assuming measurement invariance under more complicated research settings.

AUTHOR CONTRIBUTIONS

Y-YH led the implementation and manuscript writing of the study. ML designed and conducted the simulation. Both authors (Y-YH and ML) contributed to the design, analysis, interpretation of data, writing, and revising of the manuscript.

REFERENCES

Aiken, L. S., and West, S. G. (1991). *Multiple Regression: Testing and Interpreting Interactions*. Thousand Oaks, CA: Sage Publications, Inc.

Baron, R. M., and Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J. Pers. Soc. Psychol.* 51, 1173–1182.

Bohrnstedt, G. W., and Marwell, G. (1978). The reliability of products of two random variables. *Sociol. Methodol.* 9:254.

Brown, T. A. (2015). *Confirmatory Factor Analysis for Applied Research, 2nd Edn*. New York, NY: Guilford.

Busemeyer, J. R., and Jones, L. E. (1983). Analysis of multiplicative combination rules when the causal variables are measured with error. *Psychol. Bull.* 93, 549–562.

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *J. Pers. Soc. Psychol.* 95, 1005–1018. doi: 10.1037/a0013193

Chen, F. F., Sousa, K. H., and West, S. G. (2005). Teacher's corner: testing measurement invariance of second-order factor models. *Struct. Equat. Model.* 12, 471–492. doi: 10.1207/s15328007sem1203_7

Clarke, P. (2008). When can group level clustering be ignored? Multilevel models versus single-level models with sparse data. *J. Epidemiol. Community Health* 62, 752–758. doi: 10.1136/jech.2007.060798.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences, 2nd Edn*. Hillsdale, NJ: Erlbaum.

Collins, L. M., Schafer, J. L., and Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychol. Methods* 6, 330–351. doi: 10.1037/1082-989X.6.4.330

Donahue, B. H. (2006). *The Effect of Partial Measurement Invariance on Prediction*. Ph.D. thesis, University of Georgia, Athens.

Guenole, N., and Brown, A. (2014). The consequences of ignoring measurement invariance for path coefficients in structural equation models. *Front. Psychol.* 5:980. doi: 10.3389/fpsyg.2014.00980

Hox, J. J., and Maas, C. J. M. (2001). The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples. *Struct. Equat. Model.* 8, 157–174. doi: 10.1207/S15328007SEM0802_1

Hsiao, Y.-Y., Kwok, O.-M., and Lai, M. H. C. (2018). Evaluation of two methods for modeling measurement errors when testing interaction effects with observed composite scores. *Educ. Psychol. Meas.* 78, 181–202. doi: 10.1177/0013164416679877

Jaccard, J., and Wan, C. K. (1996). *LISREL Approaches to Interaction Effects in Multiple Regression*. Thousand Oaks, CA: Sage.

Jak, S., Oort, F. J., and Dolan, C. V. (2013). A test for cluster bias: detecting violations of measurement invariance across clusters in multilevel data. *Struct. Equat. Model.* 20, 265–282. doi: 10.1080/10705511.2013.769392

Judd, C. M., Kenny, D. A., and McClelland, G. H. (2001). Estimating and testing mediation and moderation in within-subject designs. *Psychol. Methods* 6, 115–134. doi: 10.1037/1082-989X.6.2.115

Judd, C. M., McClelland, G. H., and Smith, E. R. (1996). Testing treatment by covariate interactions when treatment varies within subjects. *Psychol. Methods* 1, 366–378.

Kaplan, D., and George, R. (1995). A study of the power associated with testing factor mean differences under violations of factorial invariance. *Struct. Equat. Model.* 2, 101–118.

Kim, E. S., Kwok, O.-M., and Yoon, M. (2012). Testing factorial invariance in multilevel data: a Monte Carlo study. *Struct. Equat. Model.* 19, 250–267. doi: 10.1080/10705511.2012.659623

Kim, E. S., and Willson, V. L. (2014). Measurement invariance across groups in latent growth modeling. *Struct. Equat. Model.* 21, 408–424. doi: 10.1080/10705511.2014.915374

Kreft, I. G., de Leeuw, J., and Aiken, L. S. (1995). The effect of different forms of centering in hierarchical linear models. *Multivar. Behav. Res.* 30, 1–21.

Lüdtke, O., Robitzsch, A., Kenny, D. A., and Trautwein, U. (2013). A general and flexible approach to estimating the social relations model using Bayesian methods. *Psychol. Methods* 18, 101–119. doi: 10.1037/a0029252

Maas, C. J. M., and Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology* 1, 86–92. doi: 10.1027/1614-1881.1.3.86

McNeish, D. (2017). Multilevel mediation with small samples: a cautionary note on the multilevel structural equation modeling framework. *Struct. Equat. Model.* 24, 609–625. doi: 10.1080/10705511.2017.1280797

Millsap, R. E. (1995). Measurement invariance, predictive invariance, and the duality paradox. *Multivar. Behav. Res.* 30, 577–605.

Millsap, R. E. (1997). Invariance in measurement and prediction: their relationship in the single-factor case. *Psychol. Methods* 2, 248–260.

Millsap, R. E. (1998). Group differences in regression intercepts: implications for factorial invariance. *Multivar. Behav. Res.* 33, 403–424.

Millsap, R. E. (2007). Invariance in measurement and prediction revisited. *Psychometrika* 72, 461–473. doi: 10.1007/S11336-007-9039-7

Millsap, R. E., and Kwok, O.-M. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychol. Methods* 9, 93–115. doi: 10.1037/1082-989X.9.1.93

Muthén, L. K., and Muthén, B. O. (1998–2012). *Mplus User's Guide, 7th Edn*. Los Angeles, CA: Muthén & Muthén.

Shi, D., Song, H., and Lewis, M. D. (2017). The impact of partial factorial invariance on cross-group comparisons. *Assessment* 1:1073191117711020. doi: 10.1177/1073191117711020

Vandenberg, R. J., and Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. *Organ. Res. Methods* 3, 4–70. doi: 10.1177/109442810031002

Yoon, M., and Millsap, R. E. (2007). Detecting violations of factorial invariance using data-based specification searches: a Monte Carlo study. *Struct. Equat. Model.* 14, 435–463. doi: 10.1080/10705510701301677

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

CONCEPTUAL ANALYSIS
published: 10 October 2017
doi: 10.3389/fpsyg.2017.01640

Relating Measurement Invariance, Cross-Level Invariance, and Multilevel Reliability

Suzanne Jak and Terrence D. Jorgensen*

Research Institute of Child Development and Education, University of Amsterdam, Amsterdam, Netherlands

Data often have a nested, multilevel structure, for example when data are collected from children in classrooms. This kind of data complicate the evaluation of reliability and measurement invariance, because several properties can be evaluated at both the individual level and the cluster level, as well as across levels. For example, cross-level invariance implies equal factor loadings across levels, which is needed to give latent variables at the two levels a similar interpretation. Reliability at a specific level refers to the ratio of true score variance over total variance at that level. This paper aims to shine light on the relation between reliability, cross-level invariance, and strong factorial invariance across clusters in multilevel data. Specifically, we will illustrate how strong factorial invariance across clusters implies cross-level invariance and perfect reliability at the between level in multilevel factor models.

Keywords: measurement invariance, multilevel structural equation modeling, multilevel confirmatory factor analysis, cross-level invariance, multilevel reliability

Edited by:
Holmes Finch,
Ball State University, United States

Reviewed by:
Oi-Man Kwok,
Texas A&M University, United States
Ben Kelcey,
University of Cincinnati, United States
Pietro Cipresso,
Istituto Auxologico Italiano (IRCCS),
Italy

****Correspondence:***
Suzanne Jak
s.jak@uva.nl

Specialty section:
This article was submitted to
Quantitative Psychology and
Measurement,
a section of the journal
Frontiers in Psychology

Received: *10 May 2017*
Accepted: *06 September 2017*
Published: *10 October 2017*

Citation:
Jak S and Jorgensen TD (2017)
Relating Measurement Invariance,
Cross-Level Invariance, and Multilevel
Reliability. Front. Psychol. 8:1640.
doi: 10.3389/fpsyg.2017.01640

INTRODUCTION

Multilevel data are data with a clustered structure, for instance data of children clustered in classrooms, or data of employees clustered in teams. Taking data of children in classes as an example, we can distinguish two levels in the data: we denote the child level the "within level", and the class level the "between level". Children in the same class share class-level characteristics, such as the teacher, classroom composition, and class size. Such class-level characteristics may affect child-level variables, leading to structural differences between the responses of children from different classes. With multilevel structural equation modeling (multilevel SEM), such differences are accommodated by specifying models (such as factor models) at the different levels of multilevel data. Multilevel SEM is increasingly applied in various fields such as psychology and education.

Researchers commonly interpret standardized parameter estimates, which may lead to interpretational difficulties in multilevel models. The most common standardized solution in multilevel factor models is the level-specific standardization (Hox, 2010). This type of standardization involves standardizing the within-level parameter estimates with respect to the within-level variance, and standardizing the between-level parameter estimates with respect to the between-level variance. In this standardization, it is common to find very high correlations among between-level factors, and to find standardized factor loadings that are (almost) one at the between level (e.g., Hanges and Dickson, 2006; Bakker et al., 2015). The reason that these findings are common is that residual variance at the between level is often (close to) zero (Hox, 2010), leading to relatively high standardized between-level factor loadings. At the same time, the *unstandardized* between-level factor loadings may not differ from the factor loadings at the within level. However, researchers tend to interpret the larger standardized parameter estimates at the between level as if the construct meaning is very different across the two levels of the analyses.

For example, Whitton and Fletcher (2014) found larger standardized between-level factor loadings than within-level factor loadings, and concluded that the measured construct is a "group-level construct," and that future research should emphasize interpretation at the group level rather than on the individual level. However, in the same article they reported the intraclass correlations for the subscales, showing that only 38% of the variance was at the between level, while 62% of the variance was at the individual level.

The current article explains and illustrates that neither the (near) absence of residual variance at the between level (with consequently high standardized factor loadings at the between level) nor very high reliability at the between level should be interpreted as different factors operating at the within and between level. In the next three paragraphs we briefly introduce the three concepts of measurement invariance across groups (or clusters), invariance across levels in multilevel SEM, and reliability in multilevel SEM. The goal of this article is to illuminate the relations between these three issues. Therefore, in section Relations between the three concepts we discuss each combination of concepts, and in section Example we provide illustrations with real data from students nested within schools.

Measurement Invariance Across Groups

Testing for measurement invariance is important to evaluate whether items measure the same attributes for different (groups of) respondents (Mellenbergh, 1989; Meredith, 1993). For example, if the items in a mathematical ability test measures the same attribute in boys and girls, then boys and girls with equal mathematical ability should, on average, have identical observed scores. That is, mean differences in observed scores should reflect mean differences in the true mathematical ability scores. If this is not the case, there is measurement bias. For example, given equal mathematical ability, a specific item with a worded math problem may be easier to solve for girls, because girls are generally better in reading than boys (Wei et al., 2012). For that reason, given equal levels of mathematical ability, girls might have more correct answers on this item than boys would. The item is therefore biased with respect to gender.

Structural equation modeling (SEM) with latent variables provides a flexible method to test for measurement invariance. When measurement invariance is tested with respect to a grouping variable (e.g., boys vs. girls), we can use multigroup factor analysis (MGFA) with structured means (Sörbom, 1974). In the multigroup method, specific manifestations of measurement bias can be investigated by testing across-group constraints on intercepts and factor loadings. Adequate comparisons of factor means across groups are possible if strong factorial invariance across groups holds (Meredith, 1993; Widaman and Reise, 1997). Strong factorial invariance across groups comprises equality of factor loadings and intercepts across groups. The model for the observed variables' means and covariances in group j under strong factorial invariance across groups will therefore be:

$$\boldsymbol{\mu}_j = \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\kappa}_j, \text{and} \tag{1}$$

$$\boldsymbol{\Sigma}_j = \boldsymbol{\Lambda}\boldsymbol{\Phi}_j\boldsymbol{\Lambda}' + \boldsymbol{\Theta}_j, \tag{2}$$

where $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ represent respectively the mean vector and covariance matrix of the observed variables in group j, $\boldsymbol{\kappa}_j$ and $\boldsymbol{\Phi}_j$ represent respectively the vector of common factor means and the covariance matrix of the common factors in group j, $\boldsymbol{\Theta}_j$ is the matrix with residual (co)variances of observed variables in group j, $\boldsymbol{\nu}$ is a vector of intercepts [interpretable as the means of the residual factors, Meredith and Teresi (2006)] that is invariant across groups, and $\boldsymbol{\Lambda}$ is a matrix with factor loadings (regression coefficients relating the common factor to the factor indicators) that is also invariant across groups. These equations show that if strong factorial invariance holds, differences in observed means across groups ($\boldsymbol{\mu}_j$), are a function of differences in factor means across groups ($\boldsymbol{\kappa}_j$), because nothing else on the righthand side of Equation (1) varies across groups. Also, note that the matrix with factor loadings is part of the model for the means as well as the model for the covariances. In order to provide scale and origin to the common factors, factor means and variances have to be fixed to some value in one reference group (commonly 0 for the factor means and 1 for the factor variances), and can be freely estimated in all other groups.

If the intercepts differ across groups, but the factor loadings are invariant, then strong factorial invariance is rejected, but weak factorial invariance holds. Group differences in intercepts are called "uniform bias" and differences in factor loadings are called "non-uniform bias" (Millsap and Everson, 1993).

Invariance Across Levels in Two-Level SEM

Multilevel SEM is a useful statistical technique to analyze data from many different groups, such as data from children in different school classes. Multilevel SEM then allows researchers to separate the levels of analysis (Muthén, 1990; Rabe-Hesketh et al., 2004). For example, one could evaluate differences in the students' average mathematical ability across different school classes (called the between level) and separately evaluate differences in students' relative mathematical ability within their class (called the within level). In two-level SEM, the vector of continuous response variables $\mathbf{y}_{ij}$, is split into a vector of cluster means ($\boldsymbol{\mu}_j$), and a vector of individual deviations from the respective cluster means ($\boldsymbol{\eta}_{ij} = \mathbf{y}_{ij} - \boldsymbol{\mu}_j$):

$$\mathbf{y}_{ij} = \boldsymbol{\mu}_j + \boldsymbol{\eta}_{ij}. \tag{3}$$

It is assumed that $\boldsymbol{\mu}_j$ and $\boldsymbol{\eta}_{ij}$ are independent. The covariances of $\mathbf{y}_{ij}$ ($\boldsymbol{\Sigma}_{\text{TOTAL}}$) can be written as the sum of the covariances of $\boldsymbol{\mu}_j$ ($\boldsymbol{\Sigma}_{\text{BETWEEN}}$) and the covariances of $\boldsymbol{\eta}_{ij}$ ($\boldsymbol{\Sigma}_{\text{WITHIN}}$):

$$\boldsymbol{\Sigma}_{\text{TOTAL}} = \boldsymbol{\Sigma}_{\text{BETWEEN}} + \boldsymbol{\Sigma}_{\text{WITHIN}} \tag{4}$$

The within-level and between-level covariances are modeled simultaneously but independently (unless across-level constraints are applied). For example, we may consider a two-level factor model for p observed variables and k common factors at each level:

$$\begin{aligned}\boldsymbol{\Sigma}_{\text{BETWEEN}} &= \boldsymbol{\Lambda}_{\text{BETWEEN}}\boldsymbol{\Phi}_{\text{BETWEEN}}\boldsymbol{\Lambda}'_{\text{BETWEEN}} + \boldsymbol{\Theta}_{\text{BETWEEN}},\\ \boldsymbol{\Sigma}_{\text{WITHIN}} &= \boldsymbol{\Lambda}_{\text{WITHIN}}\boldsymbol{\Phi}_{\text{WITHIN}}\boldsymbol{\Lambda}'_{\text{WITHIN}} + \boldsymbol{\Theta}_{\text{WITHIN}},\end{aligned} \tag{5}$$

where $\mathbf{\Phi}_{\text{BETWEEN}}$ and $\mathbf{\Phi}_{\text{WITHIN}}$ are $k \times k$ covariance matrices of common factors, $\mathbf{\Theta}_{\text{BETWEEN}}$ and $\mathbf{\Theta}_{\text{WITHIN}}$ are $p \times p$ (typically diagonal) matrices with residual (co)variances, and $\mathbf{\Lambda}_{\text{BETWEEN}}$ and $\mathbf{\Lambda}_{\text{WITHIN}}$ are $p \times k$ matrices with factor loadings at the between and within level, respectively.

In principle, the factor structures at the two levels can be completely different. However, in many situations the results are hard to interpret without assuming some constraints across levels. Stapleton et al. (2016) provide a nice overview of types of constructs in multilevel models. They showed that if the between-level construct represents the aggregate of the characteristics of individuals within the clusters, cross-level constraints are required. Specifically, to correctly model such constructs, the same factor structure has to apply to both levels, and factor loadings should be equal across levels. In cross-cultural research, equality of factor loadings across levels is called isomorphism (Tay et al., 2014). Across-level invariance ensures that the factors at different levels can be interpreted as the within-level and between-level components of the same latent variable (van de Vijver and Poortinga, 2002). This decomposition also allows for free estimation of the factor variance at the between level, and consequently for the calculation of the factor intraclass correlation (Mehta and Neale, 2005), representing the percentage of factor variance at the between level.

Reliability in Multilevel Factor Models

Lord and Novick (1968) defined reliability as the squared correlation between true and observed scores. An alternative (but mathematically equivalent) definition of reliability is that it is the ratio of the true score variance over the total variance (e.g., McDonald, 1999). The "true score variance" in this definition points to the part of the total score variance that is free from random error. Assuming that one has access to the true score variance, the reliability is:

$$\frac{Var(T)}{Var\,(T) + Var(E)} \tag{6}$$

where Var(T) is the true score variance, and Var(E) is measurement error variance.

In factor models, the common factor variance is used as an estimate of the true score variance. The remaining variance in an indicator stems from a residual factor ($\boldsymbol{\delta}$) that consists of two components: a reliable component, **s**, which is a stable component over persons, but not shared with other indicators; and a truly random component, **e** (Bollen, 1989). One difference between the concept of reliability in classical test theory (CTT) and the concept of reliability in the factor modeling framework is that in CTT, the variance of the stable component **s** is part of the reliable variance (included in the nominator in Equation 6), whereas in the factor analysis framework it is considered an unreliable part (only included in the denominator in Equation 6)[1].

The common factor therefore represents the reliable *common* parts of the indicators. In the SEM definition of reliability (Bollen, 1989, p. 221), the regressions of the indicator variables on the common factors represent the systematic components of the indicators, and all else represents error. The reliability of a single indicator can therefore be evaluated based on the size of the factor loading. Indices that focus on the reliability of scales with multiple indicators commonly represent some form of the ratio of common indicator variance over total indicator variance.

Geldhof et al. (2014) provided an overview of reliability estimation in multilevel factor models. They showed that level-specific reliability estimates are preferable to single-level reliability estimates when the variance at the between level is substantial. Also, they found that estimated between-cluster composite reliability (ω) was generally more unbiased than between-cluster alpha (α) and maximal reliability estimates. In this article we will therefore focus on composite reliability. Composite reliability in a congeneric factor model is defined as the ratio of *common* indicator variance over the *total* indicator variance (Werts et al., 1974; Raykov, 1997). Assuming no covariances between residual factors, and no cross loadings, composite reliability of a scale with factor variance φ, factor loadings λ_1, λ_2, ..., λ_k and residual variances θ_1, θ_2, ..., θ_k can be estimated by:

$$\omega = \frac{\left(\sum_{i=1}^{k} \lambda_i\right)^2 \varphi}{\left(\sum_{i=1}^{k} \lambda_i\right)^2 \varphi + \sum_{i=1}^{k} \theta_i} \tag{7}$$

Level-specific composite reliability is estimated by plugging in the level-specific factor loading and residual variance estimates into the formula for ω. Cluster-level reliability as estimated with Equation (7) reflects the degree to which group-level differences in a researcher's observed data can be generalized to represent between-group differences in a construct of interest (Geldhof et al., 2014).

RELATIONS BETWEEN THE THREE CONCEPTS

How Invariance between Groups Relates to between-Level Reliability

Given that in factor analysis the reliable part of the indicator is the part that reflects the common factor, reliable mean differences in observed variables between groups would reflect mean differences in common factors across groups. Lubke et al. (2003) very nicely explained the relationship between sources of within- and between-group differences and measurement invariance in the common factor model. They explicated that measurement invariance implies between-group differences cannot be due to other factors than those accounting for within-group differences.

> Suppose observed mean differences between groups are due to entirely different factors than those that account for the individual differences within a group. The notion of "different factors" as opposed to "same factors" implies that the relation of observed variables and underlying factors is different in the model for the means as compared with the model for the covariances, that is, the pattern of factor loadings is different for the two parts of the

[1] The specific variance of a measure is typically not known. In a factor model, specific variance is part of the residual variance and really only included in the denominator of Equation (6). In CTT-measures of reliability however, the specific variance may only *partly* be included in the numerator. See Bollen (1989, p. 219) for a discussion.

> model. If the loadings were the same, the factors would have the same interpretation. In terms of the multigroup model, different loadings imply that the matrix $\mathbf{\Lambda}$ in Equation (1) differs from the matrix $\mathbf{\Lambda}$ in Equation (2) (Equation numbers adjusted). However, this is not the case in the MI (measurement invariance) model. Mean differences are modeled with the same loadings as the covariances. Hence, this model is inconsistent with a situation in which between-group differences are due to entirely different factors than within-group differences (Lubke et al., 2003, p. 552).

In other words, if measurement invariance holds, then observed mean differences between groups reflect differences in the means of common factors across groups. Suppose for example that one has used several indicators to measure mathematical ability in boys and girls. Within the group of boys, the mathematical ability likely differs from boy to boy, leading to differences in the observed indicators. Similarly, within the group of girls there will be systematic differences between girls that are caused by individual differences in mathematical ability. In addition, the mean mathematical ability may differ between boys and girls. If measurement invariance holds, all group mean differences in the observed scores are caused by differences in the mean mathematical ability across groups. If the differences within and between groups are due to entirely different factors, or if there are additional factors besides mathematical ability affecting the between-group scores, then measurement invariance does not hold (Lubke et al., 2003). In this case, the measurement of between-group differences is not reliable, because differences between groups do not only reflect differences in common factors across groups.

How Invariance between Groups Relates to Invariance across Levels

When researchers are interested in differences between large numbers of groups, it becomes infeasible to conduct multigroup modeling. In these cases it is sensible to treat group as a random rather than a fixed variable, and to use multilevel techniques (Muthén and Asparouhov, 2017). For example, if a researcher wants to evaluate differences in latent variables between many countries, one could use a two-level model in which countries are treated as the clustering variable (Jak, 2017). In this example, the between-level model would represent country-level mean differences in the variables, and the within-level model would represent differences in individual deviations from the respective country means. Jak et al. (2013, 2014) provided a short overview of how three increasingly restrictive assumptions across *groups/clusters* (configural, weak, and strong factorial invariance) lead to testable restrictions across *levels* in a two-level. Specifically, they showed how weak factorial invariance across groups in a multigroup factor model translates to equal factor loadings across levels in a two-level factor model (Equations 9 and 10 in Jak et al., 2013). When strong factorial invariance holds, in addition to equal factor loadings across levels, the residual variance at the between level is zero (Equation 11 in Jak et al., 2013). We provide a more detailed and annotated derivation of these models in Appendix A in Supplementary Material. The first two columns in **Table 1** provide an overview of restrictions in a multigroup model, and the implications for a two-level model.

How Invariance Across Levels Relates to Reliability

In principle, level-specific reliability estimates can be calculated using the estimates of a two-level factor model without cross-level invariance constraints. However, in that case, the interpretation of the common factor at the two levels is not identical. In practice, research questions will often be answered using multilevel data that involves what Stapleton et al. call "configural constructs." These are constructs for which the interest is both in the within and between cluster differences, and the between-level construct represents the aggregate of the within-level characteristics. Examples are evaluation of differences in citizenship behavior within and between countries (Davidov et al., 2016) and the evaluation of teacher-student relationship quality within and between school classes (Spilt et al., 2012). These types of models require cross-level invariance restrictions on the factor loadings. When using Equation (7) to estimate composite reliability at the both levels in such a model, and provided that the two-level factor model with cross-level invariance fits the data satisfactorily, one would plug in the same unstandardized factor loadings when calculating within-level and between-level composite reliability. However, the factor variances and residual variance likely differ across levels, leading to different reliability estimates at the two levels. In the case that cluster invariance holds for all items, all residual variances at the between level will be zero, leading to perfect composite reliability at the between level (as indicated in the last column of **Table 1**). In practice, it is unlikely to find cluster invariance for *all* items, as it is unlikely that strong factorial invariance across clusters holds for *all* items. Perfect composite reliability is therefore expected to be rare in practice. Often, researchers find partial strong factorial invariance across groups (Byrne et al., 1989). Similarly, it is quite common to find

TABLE 1 | Comparison of the restrictions in a multigroup model and the implications in a two-level model with different levels of factorial invariance.

	Restrictions in multigroup model	Implications in two-level model	Implications reliability
LEVEL OF FACTORIAL INVARIANCE			
Configural	pattern($\mathbf{\Lambda}_g$) = pattern($\mathbf{\Lambda}$)	–	
Weak	$\mathbf{\Lambda}_g = \mathbf{\Lambda}$	$\mathbf{\Lambda}_{\text{WITHIN}} = \mathbf{\Lambda}_{\text{BETWEEN}}$	
Strong	$\mathbf{\Lambda}_g = \mathbf{\Lambda}, \boldsymbol{\nu}_g = \boldsymbol{\nu}$	$\mathbf{\Lambda}_{\text{WITHIN}} = \mathbf{\Lambda}_{\text{BETWEEN}}, \mathbf{\Theta}_{\text{BETWEEN}} = 0$	$\boldsymbol{\omega}_{\text{BETWEEN}} = 1$

$\boldsymbol{\nu}$ is a p-dimensional vector of intercepts. Subscript g is used for group/cluster.

TABLE 2 | Model fit of three increasingly restrictive multigroup invariance models on the well-being items.

	df	χ^2	RMSEA [90%CI]	CFI	BIC
Configural invariance	203	1742.848	0.063 [0.061; 0.066]	0.985	637061.39
Weak factorial invariance	343	3168.430	0.066 [0.064; 0.068]	0.972	636959.90
Strong factorial invariance	455	12471.471	0.118 [0.117; 0.120]	0.882	645041.28

perfect reliability for *some* of the items at the between level (e.g., Bottoni, 2016; Zee et al., 2016).

EXAMPLE

Data

We illustrate the multigroup modeling, two-level modeling, and multilevel reliability analysis using six items to measure "emotional well-being" that were included in round 2012 of the European Social Survey (Huppert et al., 2009; ESS Round 6: European Social Survey, 2014). Three items are positively formulated, asking how often in the last week a respondent was happy (WRHPP), enjoyed life (ENJLF), and felt calm and peaceful (FLTPCFL). The other three items were negatively phrased, asking how often in the last week a respondent felt depressed (FLTDP), felt sad (FLTSD), and felt anxious (FLTANX). The items were scored on a 4-point scale ranging from 0 (*none or almost none of the time*) to 3 (*all or almost all of the time*). Round 2012 of the ESS included data from 54,673 respondents from 29 countries on these items.

Analysis

All models were fit to the data with M*plus* version 7 (Muthén and Muthén, 1998–2015), using maximum likelihood estimation (MLR). This estimation method provides a test statistic that is asymptotically equivalent to the Yuan–Bentler T2 test statistic (Yuan and Bentler, 2000), and standard errors that are robust for non-normality. For illustrative purposes, we treat the responses to the 4-point scale as approximately continuous.

Statistical significance of the χ^2 statistic (using $\alpha = 0.05$) indicates that exact fit of the model has to be rejected. With large sample sizes, very small model misspecifications may lead to rejection of the model. Therefore, we also consider measures of approximate fit; the root mean square error of approximation (RMSEA; Steiger and Lind, 1980) and the comparative fit index (CFI; Bentler, 1990). RMSEA values smaller than 0.05 indicate close fit, and values smaller than .08 are considered satisfactory (Browne and Cudeck, 1992). CFI values over 0.95 indicate reasonably good fit (Hu and Bentler, 1999). In addition, for model comparison we evaluate the BIC (Raftery, 1986, 1995), of which smaller values indicate better fit.

Emotional well-being is an individual-level construct, of which the aggregated scores at the country level may differ. In the terminology of Stapleton et al. (2016), this is a configural construct, which needs cross-level equality constraints on the factor loadings.

Measurement Model

First, we fitted a two factor model to the well-being items on the merged dataset of all countries. The fit of this model was satisfactory, $\chi^2_{(8)} = 2633.591$, $p < 0.05$, RMSEA $= 0.078$, 90% CI [0.075; 0.080], CFI $= 0.98$. Inspection of modification indices showed that the modification index of a cross loading of FLTPCFL on the factor Negative well-being was around three times larger than the other modification indices. This item is the only positively phrased item that refers to feelings, while all negatively phrased items refer to feelings. Therefore, we decided to add this (negative) cross loading. The resulting model fitted the data satisfactorily, $\chi^2_{(7)} = 1352.814$, RMSEA $= 0.059$, 90% CI [0.057; 0.062], CFI $= 0.99$, and was considered the final measurement model[2].

TABLE 3 | Number of countries with a modification index of the intercept >50 and >100 per item.

	#MI > 50	#MI > 100
WRHPPY	9	4
ENJLF	13	5
FLTPCFL	10	8
FLTDPR	13	7
FLTSD	8	2
FLTANX	18	14

#MI = number of modification indices.

Multigroup Model

Next we fitted the three multigroup models representing configural invariance, weak factorial invariance, and strong factorial invariance to the data of 29 countries, with Albania as the reference country. The fit results of these three models can be seen in **Table 2**. Overall fit of the models with configural and weak factorial invariance can be considered satisfactory, but strong factorial invariance does not hold according to all fit indices. In addition, the model with weak factorial invariance has the lowest BIC-value. Apparently, at least some intercepts were not invariant across countries. Rejection of strong factorial invariance can be caused by relatively large differences in intercepts across a few countries, relatively small differences in intercepts across many countries, or a combination of both. In order to find out which items were most biased, we counted the number of countries in which each item's intercept had a high modification index. **Table 3** shows the number of countries for which specific items were flagged to be biased based on whether an intercept's modification index exceeded a threshold of 50 or

[2] The reported fit measures are obtained from an overall analysis on the merged dataset while ignoring the dependency of individuals within countries. Using an analysis with corrected fit statistic (Type = Complex in Mplus) leads to better model fit and similar conclusions, with $\chi^2_{(8)} = 327.979$, $p < 0.05$, RMSEA $= 0.027$, 90% CI [0.025; 0.030], CFI $= 0.99$ for the first model, and $\chi^2_{(7)} = 212.285$, $p < 0.05$, RMSEA $= 0.023$, 90% CI [0.021; 0.026], CFI $= 0.99$ for the modified (final) model.

TABLE 4 | Model fit of three increasingly restrictive two-level models on the well-being items.

	df	χ^2	RMSEA	CFI	BIC
Two-level CFA	14	516.692	0.026	0.976	641634.92
Cross-level invariance	19	619.519	0.024	0.972	641597.23
Strong factorial invariance	25	6880.934	0.071	0.679	647276.03

TABLE 5 | Modification indices (MIs) and chi-squared differences for releasing specific residual variances.

		free θ_i
Item	MI	$\Delta\chi^2$
WRHPPY	8895.463	661.022
ENJLF	28777.092	1229.299
FLTPCFL	40919.137	1410.159
FLTDPR	36531.309	1380.276
FLTSD	8491.897	641.51
FLTANX	147722.922	2868.184

100. Based on these counts, the item FLTANX seems to be most biased, and the item FLTSD seems the least biased.

Two-Level Model

We fitted three increasingly restrictive two-level models. The fit results can be found in **Table 4**. The first model is a two-level model specifying the measurement at the within and between levels without any constraints across levels. The fit of this model was satisfactory according to the RMSEA and CFI. However, this model does not allow for a meaningful interpretation of the factors at the two levels. Next, we constrained the factor loadings to be equal across levels, and freely estimated the factor variances at the between level. This model fitted the data significantly worse, which may be expected given the large sample size, but lead to a lower BIC-value. The overall fit was still acceptable according to the RMSEA and CFI.

Constraining the loadings to equality across levels allows computation of the factor ICC. For positive well-being, the ICC was $0.06/(1 + 0.06) = 0.057$, indicating that 5.7% of the factor variance was on the country level, and for negative well-being the ICC was $0.133/(1 + 0.133) = 0.117$, indicating that 11.7% of the factor variance was on the country level.

The model assuming strong factorial invariance, that is, the model with the between-level residual variances fixed to zero, fitted the data much worse than the first two models based on all fit indices, indicating that strong factorial invariance does not hold across countries. This finding matches the conclusion from the multigroup analysis. Non-zero residual variance at the between level shows that there are other factors than well-being influencing the country level scores on the items. **Table 5** shows the modification indices for each item's residual variance, and the actual decrease in χ^2 when freeing each item's residual variances. It is notable that, similar to the analysis of Muthén and Asparouhov (2017), the modification indices are not a good approximation of the actual drop in χ^2 when freeing the respective parameter. However, the ordering of the amount of bias present in each item is identical for the two methods. The item FLTANX seems to have the most bias, and the item FLTSD seems to be the least biased. These findings match the results from the multigroup analysis.

Figure 1 shows the unstandardized and standardized parameter estimates from the two-level model with cross-level invariance. It can be seen that although the factor loadings are constrained across levels, the standardized factor loadings are different across levels, and they are quite high at the between level, specifically for the least biased indicators. Assuming the model is configured correctly (i.e., the same construct operates at the individual and country levels), the standardized residual variance at the between level represents the proportion of item variance at the country level that is not explained by the common factor(s). These proportions are highest for the items FLTANX and FLTPCFL, and smallest for the item FLTSD, which again matches the previous conclusions about which items are most biased across countries.

Reliability

We used a two factor model with a cross loading as the measurement model. However, the formula for composite reliability that we presented (Equation 7) is only suited for congeneric factor models. Raykov and Shrout (2002) provided a method to obtain estimates of reliability for composites of measures with non-congeneric structure. Treating well-being as a multidimensional construct at each level, composite reliability for the six items was estimated as 0.77 at the within level, and 0.87 at the between level. As expected, the reliability at the between level is much higher than at the within level. The indicators that contribute most to the composite reliability estimates are the indicators with the largest standardized factor loadings (and least residual variance). For the positive well-being scale, the most reliable indicator at the between level is WRHPPY, and for the negative well-being scale the most reliable indicators are FLTDPR and FLTSD. These two items are also the items that came out as least biased in the multigroup analysis, as well as in the two-level analysis. The item with the lowest between-level standardized factor loadings is FLTPCFL, which loads on both the positive and the negative well-being factor. However, for items that load on multiple common factors, we cannot take the individual standardized factor loadings as direct indications of unbiasedness, because it does not take into account the amount of variance that is explained by the other factor(s).

DISCUSSION

The goal of our paper was to elucidate the relationship between measurement invariance across clusters, loading invariance across levels, and reliability in multilevel SEM. We used a real-data example to illustrate special issues that applied researchers should consider, which we summarize below. Invariance of loadings across levels is implied for configural constructs, so testing equality constraints on loadings across levels constitutes a test of whether a between-level construct can be interpreted as an

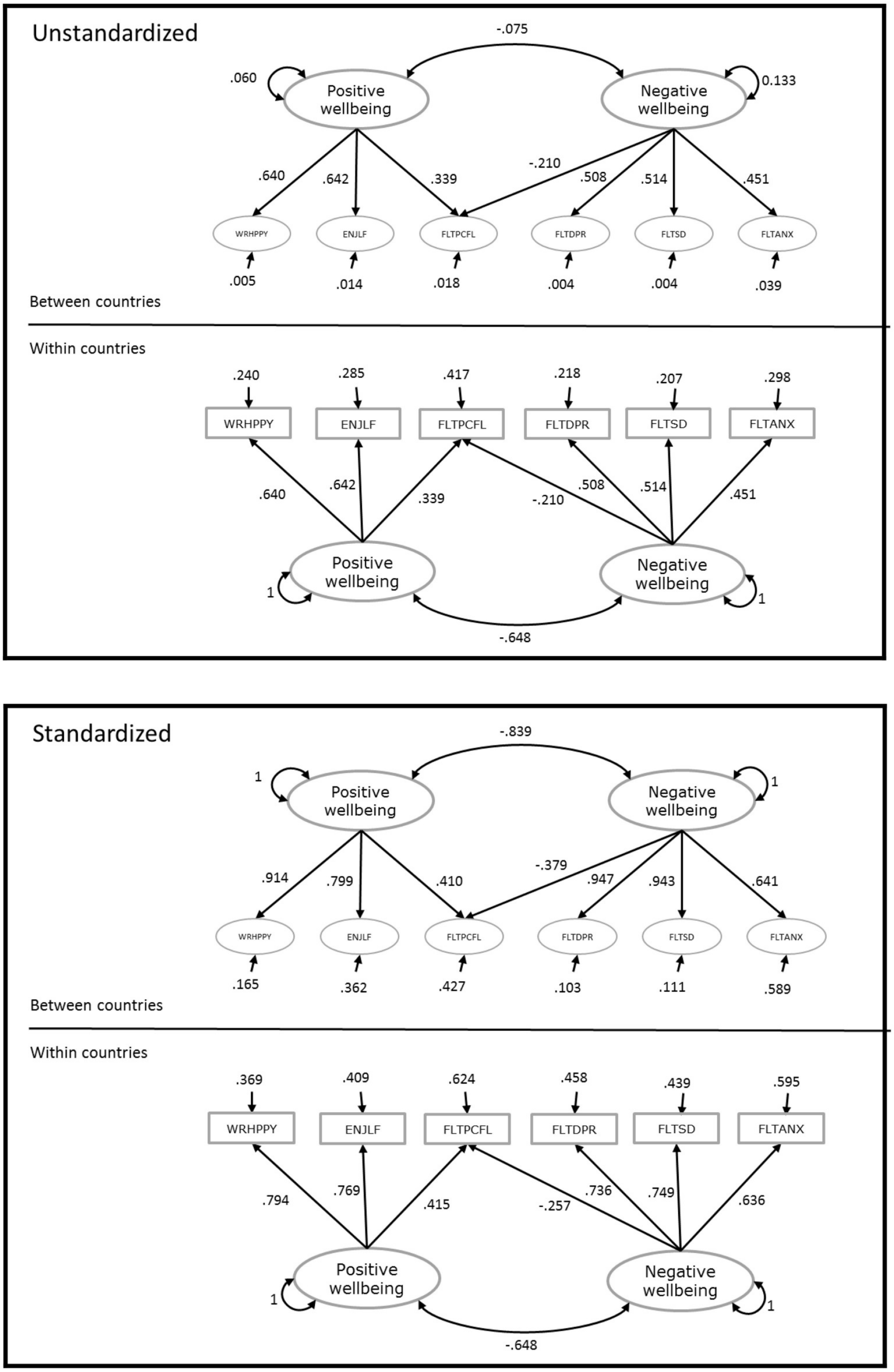


FIGURE 1 | Unstandardized and standardized parameter estimates from the two-level model with cross-level invariance.

aggregate of its within-level counterpart. Invariance of loadings across levels is also implied when factor loadings are assumed to be equal across clusters (i.e., when weak factorial invariance across clusters holds). Cross-level invariance is a necessary but not sufficient condition for weak factorial invariance across clusters. This means that if a construct cannot be regarded as configural (i.e., if cross-level invariance does not hold), then weak factorial invariance across clusters does not hold. But the reverse does not hold: If a construct *is* configural, that does not necessarily imply that weak factorial invariance across clusters also holds, because non-uniform bias across clusters has also been found to show up as residual variance at the between level (Jak et al., 2013). To summarize, equal factor loadings across clusters imply equal factor loadings across levels (and thus a configural construct), but not the other way around[3].

Equality of intercepts, on the other hand, cannot be tested across levels because the intercepts apply only to the observed variables, not separately for within- and between-level components. The common practice of fixing factor means to zero for identification of the mean structure makes it easy to show that within-level intercepts are expected to be zero. This is because the within-level component ($\boldsymbol{\eta}_{ij}$) of $\mathbf{y}_{ij}$ is partitioned from the group means ($\boldsymbol{\mu}_j$), which are the between-level components of $\mathbf{y}_{ij}$. Thus, as shown in the Appendix in Supplementary Material, the group means of $\mathbf{y}_{ij}$ are a function of $\boldsymbol{\tau}_j$ because their between-level components $\boldsymbol{\mu}_j$ are themselves a function of $\boldsymbol{\tau}_j$. Strong invariance can, however, be tested across clusters. If intercepts do not vary across clusters, that implies no between-level residual variance, so strong invariance across clusters can be tested by constraining between-level residual variances to zero in a model with cross-level loading invariance.

Finally, when working with multilevel data, reliability should be estimated separately for each level of measurement (Geldhof et al., 2014). When the construct is meant to be interpreted only at the within or between level, reliability need only be calculated at the level of interest, and a saturated model should be specified at the other level (Stapleton et al., 2016). Level-specific reliability can be interpreted for configural constructs that have analogous interpretations at each level of measurement. For example, within-level composite reliability is the proportion of variance between individuals within clusters (i.e., variability around cluster means) that is accounted for by individual differences on the within-level construct. Between-level composite reliability is the proportion of variance in cluster means that is accounted for by differences in cluster means of the same construct. Greater between-level than within-level reliability should not be mistaken for indicating that the construct has a different meaning at the between level, because (near) perfect between-level reliability (and therefore nearly zero between-level residual variance) is necessarily implied by (near) strong invariance across clusters.

AUTHOR CONTRIBUTIONS

SJ conceptualized and designed the study, SJ selected the example data and performed the analyses, TJ critically reviewed the analyses, TJ and SJ drafted the manuscript.

FUNDING

The first author was supported by a grant from The Netherlands Organization for Scientific Research: NWO-VENI-451-16-001.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2017.01640/full#supplementary-material

[3] Testing equality of factor loadings across clusters in a multilevel framework requires estimating each loading as a random slope, represented as a Level-2 factor with freely estimated variance. Testing whether a variance equals zero would constitute a test of invariance of loadings across clusters. This topic is beyond the scope of our paper but is discussed in Kim et al. (2017) and Muthén and Asparouhov (2017).

REFERENCES

Bakker, A. B., Sanz-Vergel, A. I., Rodríguez-Mu-oz, A., and Oerlemans, W. G. (2015). The state version of the recovery experience questionnaire: A multilevel confirmatory factor analysis. *Eur. J. Work Org. Psychol.* 24, 350–359. doi: 10.1080/1359432X.2014.903242

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychol. Bull.* 107, 238–246. doi: 10.1037/0033-2909.107.2.238

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. Hoboken, NJ: Wiley.

Bottoni, G. (2016). A multilevel measurement model of social cohesion. *Soc. Indic. Res.* doi: 10.1007/s11205-016-1470-7. [Epub ahead of print].

Browne, M. W., and Cudeck, R. (1992). Alternative ways of assessing model fit. *Soc. Methods Res.* 21, 230–258. doi: 10.1177/0049124192021002005

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

Davidov, E., Dülmer, H., Cieciuch, J., Kuntz, A., Seddig, D., and Schmidt, P. (2016). Explaining measurement nonequivalence using multilevel structural equation modeling the case of attitudes toward citizenship rights. *Soc. Methods Res.* doi: 10.1177/0049124116672678. [Epub ahead of print].

ESS Round 6: European Social Survey (2014). *ESS-6 2012 Documentation Report. 2.1 Edn*. Bergen: European Social Survey Data Archive, Norwegian Social Science Data Services.

Geldhof, J. G., Preacher, K. J., and Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychol. Methods* 19, 72–91. doi: 10.1037/a0032138

Hanges, P. J., and Dickson, M. W. (2006). Agitation over aggregation: clarifying the development of and the nature of the GLOBE scales. *Leadersh. Q.* 17, 522–536. doi: 10.1016/j.leaqua.2006.06.004

Hox, J. J. (2010). *Multilevel Analysis: Techniques and Applications, 2nd Edn*. New York, NY: Routledge.

Hu, L., and Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: conventional versus new alternatives. *Struct. Equat. Model.* 6, 1–55. doi: 10.1080/10705519909540118

Huppert, F. A., Marks, N., Clark, A., Siegrist, J., Stutzer, A., Vittersø, J., et al. (2009). Measuring well-being across europe: description of the ESS well-being module and preliminary findings. *Soc. Indicat. Res.* 91, 301–315. doi: 10.1007/s11205-008-9346-0

Jak, S. (2017). Testing and explaining differences in common and residual factors across many countries. *J. Cross Cult. Psychol.* 48, 75–92. doi: 10.1177/0022022116674599

Jak, S., Oort, F. J., and Dolan, C. V. (2013). A test for cluster bias: Detecting violations of measurement invariance across clusters in multilevel data. *Struct. Equat. Model.* 20, 265–282. doi: 10.1080/10705511.2013.769392

Jak, S., Oort, F. J., and Dolan, C. V. (2014). Measurement bias in multilevel data. *Struct. Equat. Model.* 21, 31–39. doi: 10.1080/10705511.2014.856694

Kim, E. S., Cao, C., Wang, Y., and Nguyen, D. T. (2017). Measurement invariance testing with many groups: a comparison of five approaches. *Struct. Equat. Model.* 24, 524–544. doi: 10.1080/10705511.2017.1304822

Lord, F. M., and Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Welsley Publishing Company.

Lubke, G. H., Dolan, C. V., Kelderman, H., and Mellenbergh, G. J. (2003). On the relationship between sources of within-and between-group differences and measurement invariance in the common factor model. *Intelligence* 31, 543–566. doi: 10.1016/S0160-2896(03)00051-5

McDonald, R. P. (1999). *Test Theory: A Unified Treatment*. Mahwah, NJ: Erlbaum.

Mehta, P. D., and Neale, M. C. (2005). People are variables too: multilevel structural equations modeling. *Psychol. Methods* 10, 259–284. doi: 10.1037/1082-989X.10.3.259

Mellenbergh, G. J. (1989). Item bias and item response theory. *Int. J. Educ. Stat.* 13, 127–143. doi: 10.1016/0883-0355(89)90002-5

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika* 58, 525–543. doi: 10.1007/BF02294825

Meredith, W., and Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Med. Care* 44, S69–S77. doi: 10.1097/01.mlr.0000245438.73837.89

Millsap, R. E., and Everson, H. T. (1993). Methodology review: statistical approaches for assessing measurement bias. *Appl. Psychol. Meas.* 17, 297–334. doi: 10.1177/014662169301700401

Muthén, B. (1990). *Mean and Covariance Structure Analysis of Hierarchical Data (UCLA Statistics Series No. 62).* Los Angeles, CA: University of California, Los Angeles.

Muthén, B., and Asparouhov, T. (2017). Recent methods for the study of measurement invariance with many groups: alignment and random effects. *Soc. Methods Res.* doi: 10.1177/0049124117701488

Muthén, L. K., and Muthén, B. O. (1998–2015). *Mplus User's Guide, 7th Edn*. Los Angeles, CA: Muthén and Muthén.

Rabe-Hesketh, S., Skrondal, A., and Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika* 69, 167–190. doi: 10.1007/BF02295939

Raftery, A. E. (1986). Choosing models for cross-classification. *Am. Sociol. Rev.* 51, 145–146. doi: 10.2307/2095483

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociol. Methodol.* 25, 111–163. doi: 10.2307/271063

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Appl. Psychol. Meas.* 21, 173–184. doi: 10.1177/01466216970212006

Raykov, T., and Shrout, P. E. (2002). Reliability of scales with general structure: point and interval estimation using a structural equation modeling approach. *Struct. Equat. Model.* 9, 195–212. doi: 10.1207/S15328007SEM0902_3

Sörbom, D. (1974). A general method for studying differences in factor means and factor structures between groups. *Br. J. Math. Stat. Psychol.* 27, 229–239. doi: 10.1111/j.2044-8317.1974.tb00543.x

Spilt, J. L., Koomen, H. M., and Jak, S. (2012). Are boys better off with male and girls with female teachers? A multilevel investigation of measurement invariance and gender match in teacher–*student relationship quality. J. School Psychol.* 50, 363–378. doi: 10.1016/j.jsp.2011.12.002

Stapleton, L. M., Yang, J. S., and Hancock, G. R. (2016). Construct meaning in multilevel settings. *J. Educ. Behav. Statist.* 41, 481–520. doi: 10.3102/1076998616646200

Steiger, J. H., and Lind, J. C. (1980). "Statistically based tests for the number of common factors," in *Paper Presented at the Annual Meeting of the Psychometric Society, Vol.* 758 (Iowa City, IA).

Tay, L., Woo, S. E., and Vermunt, J. K. (2014). A conceptual and methodological framework for psychometric isomorphism validation of multilevel construct measures. *Org. Res. Methods* 17, 77–106. doi: 10.1177/1094428113517008

van de Vijver, F. J. R., and Poortinga, Y. H. (2002). Structural equivalence in multilevel research. *J. Cross Cult. Psychol.* 33, 141–156. doi: 10.1177/0022022102033002002

Wei, W., Lu, H., Zhao, H., Chen, C., Dong, Q., and Zhou, X. (2012). Gender differences in children's arithmetic performance are accounted for by gender differences in language abilities. *Psychol. Sci.* 23, 320–330. doi: 10.1177/0956797611427168

Werts, C. E., Linn, R. L., and Jöreskog, K. G. (1974). Intraclass reliability estimates: testing structural assumptions. *Educ. Psychol. Meas.* 34, 25–33. doi: 10.1177/001316447403400104

Whitton, S. M., and Fletcher, R. B. (2014). The group environment questionnaire: a multilevel confirmatory factor analysis. *Small Group Res.* 45, 68–88. doi: 10.1177/1046496413511121

Widaman, K. F., and Reise, S. P. (1997). "Exploring the measurement invariance of psychological instruments: applications in the substance use domain," in *The Science of Prevention: Methodological Advances from Alcohol and Substance Abuse Research*, eds K. J. Bryant, M. Windle, and S. G. West (Washington, DC: American Psychological Association), 281–324.

Yuan, K. H., and Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociol. Methodol.* 30, 165–200. doi: 10.1111/0081-1750.00078

Zee, M., Koomen, H. M., Jellesma, F. C., Geerlings, J., and de Jong, P. F. (2016). Inter-and intra-individual differences in teachers' self-efficacy: a multilevel factor exploration. *J. Sch. Psychol.* 55, 39–56. doi: 10.1016/j.jsp.2015.12.003

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 02 March 2018
doi: 10.3389/fpsyg.2018.00255

Reconsidering Cluster Bias in Multilevel Data: A Monte Carlo Comparison of Free and Constrained Baseline Approaches

*Nigel Guenole**

Goldsmiths, University of London, London, United Kingdom

The test for item level cluster bias examines the improvement in model fit that results from freeing an item's between level residual variance from a baseline model with equal within and between level factor loadings and between level residual variances fixed at zero. A potential problem is that this approach may include a misspecified unrestricted model if any non-invariance is present, but the log-likelihood difference test requires that the unrestricted model is correctly specified. A free baseline approach where the unrestricted model includes only the restrictions needed for model identification should lead to better decision accuracy, but no studies have examined this yet. We ran a Monte Carlo study to investigate this issue. When the referent item is unbiased, compared to the free baseline approach, the constrained baseline approach led to similar true positive (power) rates but much higher false positive (Type I error) rates. The free baseline approach should be preferred when the referent indicator is unbiased. When the referent assumption is violated, the false positive rate was unacceptably high for both free and constrained baseline approaches, and the true positive rate was poor regardless of whether the free or constrained baseline approach was used. Neither the free or constrained baseline approach can be recommended when the referent indicator is biased. We recommend paying close attention to ensuring the referent indicator is unbiased in tests of cluster bias. All Mplus input and output files, R, and short Python scripts used to execute this simulation study are uploaded to an open access repository.

Keywords: multilevel confirmatory factor analysis, cluster bias, measurement invariance, isomorphism, homology, Monte Carlo

OPEN ACCESS

Edited by:
Suzanne Jak,
University of Amsterdam, Netherlands

Reviewed by:
Eun Sook Kim,
University of South Florida, United States
M. T. Barendse,
Ghent University, Belgium

***Correspondence:**
Nigel Guenole
n.guenole@gold.ac.uk

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *30 March 2017*
Accepted: *15 February 2018*
Published: *02 March 2018*

Citation:
Guenole N (2018) Reconsidering Cluster Bias in Multilevel Data: A Monte Carlo Comparison of Free and Constrained Baseline Approaches. Front. Psychol. 9:255. doi: 10.3389/fpsyg.2018.00255

INTRODUCTION

Measurement invariance can be demonstrated for a measurement instrument if the instrument functions equivalently, in a probabilistic sense, over subpopulations. In other words, measurement invariance exists if two individuals with equal standing on the construct being assessed, but sampled from different subpopulations, have the same expected test score. This has been explained by numerous methodologists now including Drasgow (1982, 1984), Mellenbergh (1989), Meredith (1993), Millsap (2012), and Vandenberg and Lance (2000). These authors have all shown that without demonstrating measurement invariance, conclusions about differences in latent means are dubious. More recent papers by Chen (2008) and Guenole and Brown (2014) have shown that, in addition, a lack of invariance leads to biased estimates of relationships between latent variables across groups, if the lack of invariance (or non-invariance) is not appropriately modeled.

The goal of this article is to present a Monte Carlo study that compares the effectiveness of two strategies for examining invariance simultaneously over many groups (i.e., cluster bias) in the context of multilevel confirmatory factor analysis (multilevel CFA). In the remainder of this article, we first present a brief literature review and theoretical framework for invariance testing in a multilevel CFA context. Following this overview, we describe the goals of the simulation study, the simulation conditions and rationale for their selection, and the simulation results. We then discuss the practical implications of this article for applied researchers in our discussion section.

LITERATURE REVIEW AND THEORETICAL FRAMEWORK

Methods for detecting items that violate measurement invariance are well developed. Tests for continuous indicator models are primarily based on confirmatory factor analysis (CFA). CFA tests involve first examining configural invariance, or whether the same number of latent dimensions are present in the data for each group. The configural invariance tests are followed by examining factor loadings of items across groups for equivalence. If the factor loadings are not equivalent across groups, the test items are said to violate metric or weak factorial invariance in a CFA framework.

If the factor loadings are equivalent, but intercepts are not equivalent, the items are said to violate scalar or strong invariance in CFA. Finally, equivalence of item error variances is also studied in CFA. If error variances are equal across groups, the items are said to show *strict invariance*. Chan (1998) has described the assumption of equivalent error variance as an unrealistic expectation in many applied situations. For a recent special issue on the topic of measurement invariance from the structural equation modeling perspective, see van de Schoot et al. (2015). When there are just two subpopulations of interest, say male and female in the context of CFA, measurement invariance is often examined using multiple indicator multiple causes models (Joreskog and Goldberger, 1975; Kim et al., 2012b; MIMIC: Chun et al., 2016), restricted factor analysis (Oort, 1998; RFA: Barendse et al., 2010), or multiple group models based on mean and covariance structures analysis (Sorbom, 1974; MACS: Byrne et al., 1989; Cheung and Rensvold, 1999). In these approaches the groups are fixed, and the method does not treat the groups across which invariance is examined to be a sample from a population.

Research attention has started to focus on situations where there are a large number of groups across which researchers wish to examine invariance. Examples might include invariance of a values questionnaire across cultures (e.g., Cheng et al., 2014; Cieciuch et al., 2017; Jang et al., 2017) or the invariance of a measurement instrument across classrooms in schools (e.g., Muthén, 1991). In cases like this, if there are a large number of groups, the usual multi-group approach can be cumbersome. Instead, measurement invariance can be examined with meta-analytic approaches (e.g., Cheung et al., 2006), recently developed fixed mode of variation approaches (i.e., approaches that do not attempt to make inferences beyond the groups in the analysis) like alignment optimization (Asparouhov and Muthén, 2014; Cieciuch et al., 2014), or multilevel confirmatory factor analysis (multilevel CFA: Muthén, 1994; Rabe-Hesketh et al., 2004). Multilevel CFA, the focus of this article, treats the grouping variable as a random mode of variation. In other words, it views the groups as a sample of groups from a larger population of groups (Muthén, 1994; Kim et al., 2012a; Jak et al., 2013; Ryu, 2014).

With two-level data, invariance can be examined at level-1 or level-2, but is more commonly studied at level-2, as described by Muthén et al. (1997), for example. Level-2 bias detection is the focus of the simulations to be presented in this article. Early papers by Muthén (1994) and Rabe-Hesketh et al. (2004) established the prerequisites for measurement invariance in multilevel measurement models, and a series of recent papers by Jak and her colleagues (Jak et al., 2013, 2014; Jak and Oort, 2015) further outlined the logic for tests of multilevel measurement invariance, or cluster bias. In two-level CFA the covariance matrix is decomposed as:

$$\Sigma_{\text{total}} = \Sigma_{\text{between}} + \Sigma_{\text{within}}. \tag{1}$$

If there is no measurement bias at level-2 the following models will fit the data for p observed and k latent variables:

$$\Sigma_{\text{between}} = \Lambda\Phi_{\text{between}}\Lambda' \tag{2}$$

$$\Sigma_{\text{within}} = \Lambda\Phi_{\text{within}}\Lambda' + \Theta_{\text{within}} \tag{3}$$

where $\Phi_{\text{ between}}$ and $\Phi_{\text{ within}}$ are $k \times k$ latent variable covariance matrices, Λ is a $p \times k$ matrix of factor loadings, and $\Theta_{\text{ within}}$ is a diagonal $p \times p$ matrix of residual variances. Cluster bias is related to the concept of isomorphism, which refers to equal factor loadings across levels. Isomorphism has important consequences for conclusions about the similarity of relationships between variables across levels, which in turn is referred to as homology (Tay et al., 2014; Guenole, 2016). However, absence of cluster bias is a stronger assumption than isomorphism, because cluster bias refers to non-zero residual variance at level-2, in a model where isomorphism holds.

In practice, the Jak et al. (2013) procedure for testing multilevel invariance unfolds as follows. First, *configural invariance* is examined. The configural invariance model holds where the pattern of factor loading coefficients is consistent across the within and between levels of the multilevel CFA model. Next, the *cluster invariance* model is fit to the data where factor loadings and intercepts are equivalent across clusters. The data support a *cluster invariance* or strong invariance model when the factor loadings are equivalent across levels and level-2 item residual variances are not significantly different from zero. Jak and Oort (2015) noted that the result of bias in factor loadings and intercepts manifests as level-2 residual variance when factor loadings are constrained across levels, and the test for cluster bias does not differentiate the source for the bias (i.e., whether it is intercept or factor loading non-invariance across level-2 clusters), rather, it simply tests for the presence of measurement bias. If the level-2 residual variances are significantly different from zero, the bias could be in factor loadings and/or intercepts. In this article, we focus attention on uniform bias which we simulate by incorporating the direct effect of a level-2 violator, described more in the methodology section.

In the sections that follow, we contrast free and constrained baseline approaches to testing measurement bias. Importantly, the free vs. constrained distinction in the context of item level testing is distinct from scale level testing for cluster bias where the configural model is contrasted with the scalar invariance model. The item level procedures we investigate here are likely to be followed by researchers if they find that cluster invariance is violated.

Constrained Baseline Approaches to Measurement Invariance Testing

An item level approach to cluster bias based on Jak et al.'s (2013) procedure can be considered a constrained (of "fixed") baseline approach. This is because it begins by fixing all parameters to be tested to be either equal across levels in the case of factor loadings or zero in the case of between level residual variances. In this approach, the overall fit of the fully constrained model is first evaluated. The alternative model then frees the level-2 residual variances for the studied item(s). An evaluation of the improvement in model fit from freeing the item parameters is then made, using a test such as the likelihood ratio difference test or one of its variations. Statistical significance indicates that the model with constraints fits significantly worse than the model without constraints, and the item exhibits measurement bias. Conversely, statistical non-significance indicates that the model with constraints does not fit significantly worse than the model without constraints, and the item does not exhibit measurement bias. Stark et al. (2006) noted that constrained baseline procedures are the typical approach used by researchers coming from an item response modeling tradition.

Free Baseline Approaches to Measurement Invariance Testing

An alternative approach to the constrained baseline strategy begins with a free baseline where minimal constraints are imposed. In this approach, the minimally identifiable model is estimated as the baseline model. The alternative model then constrains the parameters of the items being tested across groups (one factor loading and one residual variance), but leaves parameters for all other items free across groups. An evaluation of the difference in fit between the constrained and unconstrained models is then made using a test such as the likelihood ratio test. Statistical non-significance indicates that fixing the item parameters equal across groups does not yield a statistically significant decrement in model fit, and that the item does not exhibit bias. Statistical significance indicates that the item does exhibit measurement bias across groups. Stark et al. (2006) noted that free baseline procedures are the typical approach used by researchers coming from a structural equation modeling tradition.

Competing Rationales for Constrained and Free Baseline Approaches

Readers should note that while the free vs. constrained baseline issue has not been examined in the context of multilevel measurement invariance, numerous related examples exist in the traditional two-group case under the label of *iterative* bias detection. These include applied instances (e.g., Navas-Ara and Gómez-Benito, 2002) as well as Monte Carlo studies (e.g., Oort, 1998; Barendse et al., 2012). In the iterative approach, results of previous item tests are incorporated into the baseline model for testing of subsequent items. In the method adopted in this article, we always revert to the original free or constrained baseline for tests of subsequent items.

Researchers have offered different rationales for examining invariance with free and constrained baseline approaches. Constrained baseline approaches might be a reasonable approach if the majority of items are believed to be invariant, perhaps based on past research. Furthermore, Jak and Oort (2015) showed with simulations that a constrained baseline approach leads to reasonably accurate conclusions under some circumstances when examining cluster bias. A constrained baseline might also be defended on the basis that more stable parameter estimates are achieved when the linking required to establish a common metric is based on more than one item (Stark et al., 2006).

Nevertheless, there may be an impact on subsequent tests of model fit if the unrestricted model is misspecified. To calculate the log-likelihood ratio test two models are estimated, an unrestricted model M_0, and a restricted model, M_1. The log-likelihood statistic is calculated by comparing the log-likelihoods of the two models: LRT $= -2$ x $(\ell_0 - \ell_1)$. The LRT statistic is distributed as χ^2 with degrees of freedom equal to the difference in the number of estimated parameters between the models, but this is only so if the unrestricted model is correctly specified. If it is not, it could see a reduction in statistical power to detect bias when it exists, and increased Type I errors where non-biased items are identified as biased. From a logical perspective, the cautious and strongest theoretical approach seems to be to use the free baseline procedure. Indeed, simulation studies in the two-group context have revealed greater accuracy for the free baseline approach across dominance and unfolding item response models (Stark et al., 2006; Wang et al., 2013; Chun et al., 2016). Comparing the free and constrained baseline approaches with Monte Carlo methods in a multi-level CFA context is the central goal of this article. Importantly, while the general procedures of free and constrained approaches to invariance testing are not new, the two procedures we examine have never been evaluated before in the context of testing cluster bias. There are also other possible approaches to free and constrained baseline testing that this article does not address. We return to these alternative approaches in our discussion.

We broadly follow the recommendations of Paxton et al. (2001). Our hypotheses were as follows:

H1. A free baseline approach will provide greater decision accuracy in terms of true positive rates and false positive rates in comparison to the constrained baseline approach.

H2. The improved performance of the free baseline approach will be even more observable in terms of true positive rates and false positive rates with factors expected to increase decision accuracy (i.e., higher ICC, larger L1 and L2 sample sizes, more non-invariant items, and higher magnitude bias).

We were interested in the performance of free and constrained baseline approaches where the referent item was free of bias and when it was biased, since in practice this important assumption may be easily violated. We did not have a hypothesis for the biased referent indicator section in our design and therefore treat this part of the analysis as exploratory.

METHOD

Design Features

The simulation conditions for the most part follow the set up described by Jak et al. (2014), which provide a strong basis on which to evaluate the performance of cluster bias detection in multilevel CFA. In addition, we verified the appropriateness of these conditions by a review of existing simulation studies addressing multilevel CFA questions. In the sections that follow, we describe the simulation set up for the current study.

Fixed Features of Simulation Design

Test length was set at five items. Five items have been commonly used in measurement model simulation studies, and this number of items is very common in survey research where there is not sufficient room for longer scales.

Continuous indicators were simulated. Continuous item indicator factor models are common in survey work where research shows that so long as the number of scale points in a Likert scale model is greater than five, continuous factor models perform well.

Replications were set at 500 replications per cell which is consistent with the number of replications used in past studies and is expected to be a sufficient number of replications to achieve reliable results.

Missing data patterns were not included in our simulation, and so the impact of missing data patterns in multilevel measurement invariance falls outside the scope of our simulation.

Level-2 violators were simulated to introduce non-invariance in our simulations. The effect of the level-2 violator is to increase the variance for the biased item. We subsequently examined items for bias with tests of cluster bias.

Experimental Conditions

Level 1 Samples Sizes (Three Levels)

Level-1 sample sizes (L1N) were set at 2, 5, and 25, mirroring the simulation conditions presented since cluster samples sizes of two are common in dyad research, five are common in small group research, and 25 is common in educational and organizational research.

Level 2 Sample Sizes (Two Levels)

Level-2 sample sizes (L2N) were set at 50 or 100. These cluster sizes can be considered moderate and large, and were chosen because results of simulations in multilevel CFA contexts by Maas and Hox (2005) show sample sizes in this range are required for estimation accuracy.

Intraclass Correlations (ICCs) (Three Levels)

Based on a review of ICCs in simulation studies including Maas and Hox (2005), the ICCs were set 0.10, 0.20, and 0.30. While smaller ICCs have been investigated by some researchers (e.g., Depaoli and Clifton examined ICCs of 0.02), larger ICCs are both common in applied settings and less likely to result in inadmissible solutions.

Number of Biased Items (Three Levels)

We included conditions with 0, 1, and 2 biased items. This allowed examining the impact of the severity of baseline misspecification on decision accuracy. The no bias condition was included simply as a simulation baseline to check the basal Type I error, following Kim et al. (2012b).

Size of Bias (Three Levels)

We incorporated three levels of bias: no bias, small bias, and large bias. We set the bias to be one and five percent of the total variance of the indicator for small and large bias respectively. We did this using the methodology of increasing the variance of the biased item by allowing a direct effect of a level-2 violator variable, described below under data generation. This approach has been used in past simulation studies, including Barendse et al. (2012). Again, the no bias condition was included only as a simulation baseline check.

Biased or Unbiased Anchor Item (Two Levels)

We examined the performance of the free and constrained baseline models when the anchor item was biased and when it was unbiased. Note that for levels of bias with 2 biased items under the biased anchor item this meant there were, in fact, 3 biased items: 2 biased items with a biased anchor item. The size of the bias in the anchor item was set to match the size of the item bias in the remaining item(s), i.e., 1 or 5% bias.

Summary of Factorial Design

Our design included the baseline check conditions of three L1N × two L2N × three intraclass correlation levels x one combination of biased items and levels of bias (i.e., zero bias) = 18 conditions; along with three L1N × two L2N × three intraclass correlation levels x four combinations of bias items and levels of bias (i.e., one or two bias items with either small, or large bias) x two anchor item levels (biased and unbiased) = 144 conditions. Each condition was analyzed using two detection strategies (i.e., both the free and the constrained baseline strategies).

Data Generation

All simulated data were generated as continuous multivariate normal using Mplus version 8 (Muthén and Muthén, 1998-2017). All input scripts, outputs scripts, and python scripts used to extract model summary statistics are available at the following figshare link: https://figshare.com/s/23427e33be46d406b5d0. The base model from which all simulation models can be derived is presented in **Figure 1**. In the unbiased referent indicator

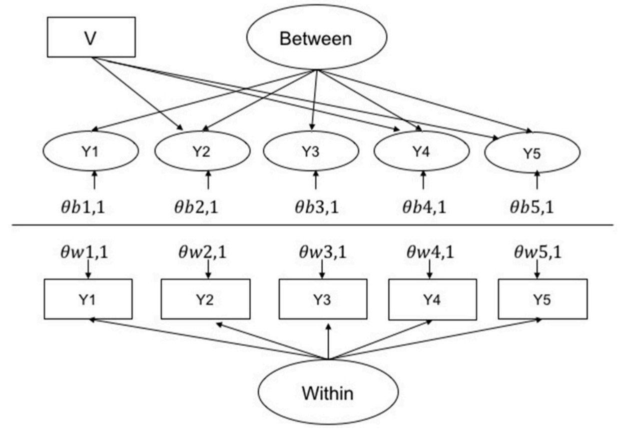


FIGURE 1 | Multilevel CFA base model for the most biased condition (biased referent 2, and two further biased items 4 and 5). V is a level 2 violator that causes bias in the affected indicators by increasing there variance by either 1 and 5%. The variance of the within factor was set at 1 and the between was varied to adjust the ICC. Loading on within and between were set at 1 for all items. θs are errors where w and b represent within and between, respectively, and 1, 1 indicates the first indicator of the factor, etc.

conditions, bias was simulated on items 2 and 4. In the biased referent indicator conditions, bias was simulated on items 2, 4, and 5.

Model Identification

Unbiased Anchor

To identify the metric of the latent factors in the free baseline approach we fixed the first factor loading of each factor on within and between levels at one. This item was not biased. The baseline model in the constrained approach was identified by fixing the factor loading of the first indicator of each item at 1 and having all remaining within and between level factor loadings equal and all level two residual variances set at zero—thus violating the assumption of an unbiased referent indicator.

Biased Anchor

To identify the metric of the factors in the free baseline approach we fixed the second factor loading of each factor on within and between levels at one. This item was a biased item. All remaining factor loadings and residual variances were freely estimated. To identify the metric of the latent factors in the constrained baseline condition, we again fixed the within and between level factor loadings of item 2, which was biased, at 1. All remaining item factor loadings were constrained equal across levels and all level-2 residual variances were fixed at 0.

Estimation

We estimated all models using robust maximum likelihood estimation (MLR).

Testing for Cluster Bias From Constrained and Unconstrained Baselines

To test the invariance of items in the free baseline condition, we examined the significance of the difference in $-2 \times$ the log-likelihood of the restricted model and the unrestricted model. In the restricted model the within and between level factor loadings were equal and the residual variances for the tested item was zero, respectively. In the unrestricted model the factor loadings and residual variance for the tested item were free. This yields a test statistic that is χ^2 distributed with 2 degrees of freedom. To test for bias of items under the constrained baseline, the starting model was the reduced model where all factor loadings were fixed equal and level-2 residual variances were zero; the unrestricted model freed the within and between level factor loadings for the tested item along with its between level residual variances. The test statistic was compared against a χ^2 distribution with two degrees of freedom.

This restricted baseline approach differs from the Jak et al. (2014) procedure in that their approach evaluates the improvement in fit from moving from the model with constrained factor loadings for an item with its residual variance at zero to the same model but with only the residual variance freed. However, for comparability with the free baseline approach outlined, here we examine the improvement in fit that results from freeing both the item residual variance and the across level factor loading constraint simultaneously.

A correction is sometimes applied to calculate the differences in χ^2 between nested models because differences in $-2 \times$ log-likelihood values are not χ^2 distributed under the maximum likelihood estimator with robust standard errors (Satorra and Bentler, 2001). However, Jak and Oort (2015, p. 440), citing Cham et al. (2012) recommended using the uncorrected difference in $-2 \times$ log-likelihood between the nested models because in the context of difference testing a procedure with the correction does not perform better than an approach without the correction. They reported that their log-likelihood differences were sometimes negative, and they considered these non-significant. A strictly positive χ^2 difference test has been developed for the situations where the negative values occur, and it has been suggested as potentially being relevant in other cases such as small samples. However, recommendations to date are unclear with regard to whether this strictly positive variation ought to be applied in all cases. For this reason, we used unscaled $-2 \times$ log-likelihood difference tests.

True Positive Rates and False Negative Rates

True positives rates were calculated as the proportion of simulation runs within each condition where biased items were correctly identified as biased with the two degree of freedom log-likelihood difference test. False positive rates were calculated as the proportion of simulation runs within each condition where unbiased items were incorrectly identified as exhibiting bias with the two degree of freedom log-likelihood difference test. Power corresponds to the true positive rate, while Type I error corresponds to the false positive rate.

When estimating many multilevel models with small variances and low L1N, numerous estimation challenges are likely to emerge, particularly in these models. These include runs where (a) the software does not complete a replication, (b) the software makes estimation adjustments due to the estimation hitting saddle points, (c) models converge to inadmissible solutions, and (d) negative log-likelihood difference values result. In instances when either model required for a log-likelihood difference did not converge, the result was not counted as a true positive or a false positive, but the proportion of true positives and false positives observed for the cell is still expressed as a proportion of the 500 intended runs.

We observe differences in reporting practices with regard to whether runs with adjustments due to saddle points and inadmissible solutions are summarized over or omitted in Monte Carlo results for measurement invariance. In this study, if the software converged to an inadmissible solution, and when an adjustment was made due to hitting a saddle point, the log-likelihood difference test was still conducted and summarized in the same way as for log-likelihood difference tests for admissible solutions. In addition, as with the study reported by Jak et al. (2013), on numerous occasions the log-likelihood difference was negative. In the Jak et al. (2013) study, the authors counted these to be non-significant differences, however, in the present study we considered these to be inconclusive and did not count them in our analyses as constituting a true positive or a false positive occurrence.

RESULTS

Simulation Baseline Check

Under the constrained baseline approach with no biased items the false positive rate (based on testing item 2 for bias) and the true positive rate (based on testing item 3 for bias) are both false positive rates, because there is no bias. The detection rate should be around the nominal significance level of 0.05 across all experimental conditions, because the baseline model and the comparison models are always correctly specified. **Table 1** reveals that the false positive rate based on testing item 2 for bias in the constrained baseline condition was always slightly lower than 0.05. The false positive rate based on testing item 3 for bias was also slightly lower than 0.05 across all conditions. Similarly, under the free baseline approach, the false positive rate (based on testing item 2 for bias) and the true positive rate (based on testing item 3 for bias) both constitute false positive rates. These are expected to be around 0.05 across all experimental conditions. **Table 1** shows that the false positive rate based on testing item 2 for bias was always slightly lower than 0.05, and the false positive rate based on testing item 3 for bias was also always slightly lower than the expected 0.05.

Negative Log-Likelihood Difference Tests

In the simulation baseline conditions, and in all conditions that follow, the occurrences of negative log-likelihood difference tests were substantially higher under the constrained baseline detection strategy, and precise frequencies can be observed in **Table 2** through **Table 5**.

Unbiased Referent Item Results

Summary of True Positive and False Positive Rates

True positive (power) and false positive (Type I error) rates are summarized in **Tables 2**, **3**. For the unbiased anchor conditions, the overall true positive rate for the free baseline approach was 0.44, while the overall true positive rate for the constrained baseline approach was similar at 0.42. The overall false positive rate for the free baseline approach was 0.04, while the overall false positive rate for the constrained baseline was unacceptably high at 0.14. This is because in contrast to the baseline check condition where the baseline model was always correctly specified and the fixed baseline approach performed well in terms of false positives, in these conditions the baseline model was always misspecified and the false positives are too high. We now further explore factors associated with variability in these true positive and false positive rates with ANOVA models.

True Positive Rates (Power)

In an ANOVA model for the unbiased anchor condition where the true positive rate was predicted by all independent variables the significant independent variables at $p < 0.05$ were Level-2 N ($\eta^2_{\text{L2N}} = 0.031$), Level-1 N ($\eta^2_{\text{L1N}} = 0.566$), the number of biased items ($\eta^2_{\text{no. biased items}} = 0.011$) and the size of the bias ($\eta^2_{\text{size bias}} = 0.175$). Non-significant effects included ICC ($\eta^2_{\text{ICC}} = 0.003$), and free vs. fixed ($\eta^2_{\text{free v fixed}} = 0.001$, where the free baseline was coded 0 and the constrained baseline was coded 1). It is important not to over-interpret small but statistically significant effects, so here we consider effect sizes in relation to Cohen (1988) criteria. The only effects that met Cohen's (1988) criterion for being at least a moderate effect (0.058) were Level-1 N and the size of bias. Next, an examination of all two-way interactions indicated that the interaction between the ICC and the number of biased items was significant at $p < 0.05$ ($\eta^2_{\text{ICC} \times \text{no. biased items}} = 0.008$), although by Cohen's benchmark this effect is small. There were no interactions involving the free vs. fixed baseline variable manipulation.

False Positive (Type I Error) Rates

In an ANOVA model for the unbiased anchor condition where the false positive rate was predicted by all independent variables the significant independent variables at $p <0.05$ were L1N ($\eta^2_{\text{L1N}} = 0.097$), number of biased items ($\eta^2_{\text{no. biased items}} = 0.071$), size of bias ($\eta^2_{\text{size bias}} = 0.036$), and whether a free or fixed baseline was used ($\eta^2_{\text{free v fixed}} = 0.072$). These effects are all moderate by Cohen's (1988) criteria, aside from the effect of the size of the bias, which was small. We next explored all two-way interactions. This indicated that the following interactions were significant at $p < 0.05$: L1N × number of biased items ($\eta^2_{\text{L1N} \times \text{no. biased items}} = 0.067$), L1N × size of bias ($\eta^2_{\text{L1N} \times \text{size bias}} = 0.072$), L1N × free vs. fixed baseline ($\eta^2_{\text{L1N} \times \text{free v fixed}} = 0.100$), number of biased items × size of bias ($\eta^2_{\text{no. biased items} \times \text{size bias}} = 0.022$), number of biased items × free vs. fixed baseline ($\eta^2_{\text{no. biased items} \times \text{free v fixed}} = 0.055$), and size of bias × free vs. fixed baseline ($\eta^2_{\text{size bias} \times \text{free v fixed}} =0.032$). Interaction effect sizes that were at least moderate by Cohen's standard, therefore, include L1N × number of biased items, L1N × size of bias, L1N × free vs. fixed baseline ($\eta^2_{\text{L1N} \times \text{free v fixed}} = 0.100$). The interaction for the number of biased items × free vs. fixed baseline was also very close to being a moderate effect size.

TABLE 1 | True positive rates and false positive rates for unbiased anchor with no bias.

						Free baseline				Fixed baseline			
Cell	**L2N**	**L1N**	**ICC**	**Items**	**Size (%)**	**NLD**	**FP-2**	**NLD**	**FP-3**	**NLD**	**FP-2**	**NLD**	**FP-3**
1	50	2	0.10	0	0	21	0.03	13	0.02	225	0.02	229	0.02
31	50	2	0.20	0	0	20	0.02	18	0.03	243	0.02	226	0.02
61	50	2	0.30	0	0	17	0.03	10	0.03	232	0.02	237	0.02
6	50	5	0.10	0	0	49	0.02	21	0.02	253	0.01	239	0.01
36	50	5	0.20	0	0	20	0.03	31	0.02	193	0.02	213	0.03
66	50	5	0.30	0	0	24	0.01	27	0.02	211	0.01	215	0.02
11	50	25	0.10	0	0	38	0.02	34	0.02	130	0.01	121	0.03
41	50	25	0.20	0	0	33	0.03	28	0.01	105	0.04	122	0.01
71	50	25	0.30	0	0	32	0.03	23	0.02	109	0.03	104	0.01
16	100	2	0.10	0	0	26	0.03	26	0.03	277	0.01	268	0.02
46	100	2	0.20	0	0	37	0.04	19	0.03	248	0.02	258	0.02
76	100	2	0.30	0	0	36	0.03	41	0.02	215	0.02	241	0.01
21	100	5	0.10	0	0	20	0.03	29	0.04	186	0.02	208	0.03
51	100	5	0.20	0	0	22	0.02	24	0.02	202	0.02	172	0.02
81	100	5	0.30	0	0	22	0.02	34	0.03	214	0.02	203	0.03
26	100	25	0.10	0	0	31	0.02	39	0.02	107	0.01	126	0.02
56	100	25	0.20	0	0	32	0.02	34	0.03	108	0.03	127	0.03
86	100	25	0.30	0	0	26	0.02	32	0.03	102	0.02	123	0.02

L2N, level-2 sample size; L1N, level –1 sample size; ICC, intraclass correlation; Items, number of biased items; NLD, count of negative log-likelihood difference test results; FP-2, false positive rate for item 2; FP-3, false positive rate for item 3.

We explored the interactions involving the free vs. fixed baseline manipulation, our focal independent variable, further with graphical plots. **Figure 2** depicts the interaction between L1N and free vs. fixed baseline on the false positive rate. This figure reveals that moving from the lowest L1N size to the highest L1N size for the free baseline model results in no notable change in the false positive rate. On the other hand, moving from the lowest L1N size to the highest L1N size with a fixed baseline leads to a substantial jump in the false positive rate. **Figure 3** depicts the interaction between the number of biased items and the free vs. fixed baseline strategy. This figure reveals that with 1 biased item present, the free and fixed baseline approaches perform similarly in terms of controlling the Type I error rate.

However, in the presence of two biased items, the free baseline approach performs considerably better at controlling the Type I error rate. The final interaction involving our focal independent variable was for free vs. fixed and the size of the bias. This interaction is plotted in **Figure 4**. It reveals that when the size of the bias is small, the free and fixed baseline approaches perform similarly, albeit with a lower Type I error rate for the free baseline approach. When the size of the bias is large, the free baseline approach continues to control the Type I error rate appropriately. The Type I error rate for the fixed baseline approach, however, rises to an unacceptable level.

Biased Referent Item Results

Summary of True Positive and False Positive Rates

True positive (power) and false positive (Type I error) rates for the biased referent indicator are summarized in **Tables 4**, **5**. These tables indicate that the false positive rate is poorly controlled when the anchor item is biased, regardless of whether a free baseline or a fixed baseline approach is adopted. The overall false positive rates under the free and fixed baseline approach were both unacceptable when the anchor item was biased, at 0.25 and 0.29 respectively. In this biased anchor condition, the overall true positive rate under the free baseline approach was 0.16, while it was higher at 0.28 under the fixed baseline approach. Comparison of the relative advantages of the free and fixed baseline across conditions is not meaningful given the unacceptably high false positive rates and poor power across both approaches. We do not explore this issue further here, instead we return to the topic of identifying an unbiased item in the discussion section below.

DISCUSSION

Breakthroughs in measurement invariance methods have made techniques available for testing measurement invariance across high numbers of groups with relatively small within group sample sizes. This is an important development, because until now the idea of testing whether different groups interpret survey questions similarly has been limited to a small number of groups with large sample sizes. Yet, the failure to adequately establish a common interpretation across groups is known to cause problems for interpretations of differences in latent means and relationships between latent variables. The method studied in this article to test measurement invariance, continuous indicator multilevel confirmatory factor analysis, is ideal for studying measurement invariance (i.e., cluster bias) across many groups.

TABLE 2 | True positive and false positive rates for unbiased anchor with one biased item.

						Free baseline				Fixed baseline			
Cell	**L2N**	**L1N**	**ICC**	**Items**	**Size (%)**	**NLD**	**TP**	**NLD**	**FP**	**NLD**	**TP**	**NLD**	**FP**
2	50	2	0.10	1	1	12	0.04	19	0.04	207	0.02	200	0.03
3	50	2	0.10	1	5	4	0.14	21	0.06	216	0.01	216	0.01
32	50	2	0.20	1	1	25	0.05	15	0.05	198	0.02	220	0.02
33	50	2	0.20	1	5	7	0.09	20	0.04	84	0.07	253	0.02
62	50	2	0.30	1	1	18	0.06	17	0.05	208	0.04	221	0.02
63	50	2	0.30	1	5	2	0.13	27	0.04	75	0.12	240	0.02
7	50	5	0.10	1	1	11	0.07	20	0.02	137	0.07	225	0.02
8	50	5	0.10	1	5	1	0.47	16	0.04	11	0.51	169	0.04
37	50	5	0.20	1	1	13	0.06	20	0.02	130	0.07	188	0.02
38	50	5	0.20	1	5	1	0.59	21	0.02	5	0.60	193	0.02
67	50	5	0.30	1	1	12	0.06	36	0.02	98	0.08	189	0.03
68	50	5	0.30	1	5	0	0.66	15	0.02	3	0.69	198	0.03
12	50	25	0.10	1	1	2	0.57	25	0.01	4	0.60	99	0.03
13	50	25	0.10	1	5	0	1.00	16	0.02	0	1.00	56	0.07
42	50	25	0.20	1	1	0	0.68	21	0.02	4	0.71	84	0.02
43	50	25	0.20	1	5	0	1.00	21	0.02	0	1.00	59	0.08
72	50	25	0.30	1	1	1	0.75	33	0.02	1	0.78	111	0.03
73	50	25	0.30	1	5	0	1.00	26	0.02	0	1.00	59	0.10
17	100	2	0.10	1	1	21	0.07	27	0.06	203	0.02	244	0.01
18	100	2	0.10	1	5	10	0.18	28	0.04	72	0.16	254	0.02
47	100	2	0.20	1	1	18	0.03	31	0.04	190	0.03	245	0.03
48	100	2	0.20	1	5	5	0.20	30	0.04	44	0.18	227	0.02
77	100	2	0.30	1	1	17	0.03	20	0.04	209	0.02	212	0.03
78	100	2	0.30	1	5	8	0.23	39	0.01	47	0.24	235	0.02
22	100	5	0.10	1	1	15	0.09	24	0.03	102	0.10	170	0.01
23	100	5	0.10	1	5	0	0.84	27	0.03	0	0.87	175	0.04
52	100	5	0.20	1	1	7	0.11	23	0.02	64	0.10	186	0.02
53	100	5	0.20	1	5	0	0.89	25	0.03	0	0.92	170	0.03
82	100	5	0.30	1	1	8	0.10	20	0.03	67	0.10	197	0.02
83	100	5	0.30	1	5	0	0.94	30	0.04	0	0.96	146	0.04
27	100	25	0.10	1	1	1	0.87	28	0.04	1	0.89	91	0.04
28	100	25	0.10	1	5	0	1.00	31	0.02	0	1.00	36	0.15
57	100	25	0.20	1	1	0	0.94	29	0.02	0	0.95	97	0.05
58	100	25	0.20	1	5	0	1.00	23	0.02	0	1.00	40	0.18
87	100	25	0.30	1	1	0	0.98	30	0.03	0	0.99	86	0.04
88	100	25	0.30	1	5	0	1.00	29	0.02	0	1.00	37	0.18

L2N, level-2 sample size; L1N, level −1 sample size; ICC, intraclass correlation; Items, number of biased items; NLD, count of negative log-likelihood difference test results; TP, true positive rate; FP, false positive rate.

So far, it has been implemented using a constrained baseline approach, where the starting model has all factor loadings equal across levels and all level-2 residual variances fixed at 0. However, the growing literature on free baseline approaches suggests that a free baseline approach might have greater decision accuracy for bias detection. This article examined whether this is also the case for multilevel confirmatory factor analysis tests of measurement invariance. Indeed, support for a free baseline approach in a multilevel CFA setting was observed.

Overall, the power for the free baseline approach when the referent indicator was unbiased was 0.44. This was similar, albeit slightly higher, than the power for the constrained baseline approach under these conditions at 0.42. The real difference between the two methods when the referent indicator was unbiased was observed in the false positive rates. The overall false positive rate for the unbiased referent indicator under the free baseline was 0.04, which is acceptable. The false positive rate for the constrained baseline approach was unacceptably high at 0.14. When the referent indicator is unbiased, the free baseline approach should be preferred. Our first hypothesis, that a free (as opposed to a constrained or “fixed”) baseline approach would have an accuracy advantage in terms of true positive

TABLE 3 | True positive and false positive rates for unbiased anchor with two biased items.

						Free baseline				Fixed baseline			
Cell	**L2N**	**L1N**	**ICC**	**Items**	**Size (%)**	**NLD**	**FP**	**NLD**	**FP**	**NLD**	**FP**	**NLD**	**FP**
4	50	2	0.10	2	1	12	0.05	13	0.03	206	0.03	239	0.01
5	50	2	0.10	2	5	5	0.16	21	0.04	186	0.06	193	0.02
34	50	2	0.20	2	1	25	0.05	14	0.04	235	0.03	227	0.02
35	50	2	0.20	2	5	10	0.19	14	0.04	157	0.08	181	0.03
64	50	2	0.30	2	1	13	0.05	20	0.03	197	0.04	193	0.03
65	50	2	0.30	2	5	8	0.18	13	0.06	115	0.06	174	0.04
9	50	5	0.10	2	1	14	0.05	21	0.03	159	0.03	194	0.02
10	50	5	0.10	2	5	4	0.46	17	0.02	42	0.23	113	0.07
39	50	5	0.20	2	1	11	0.05	19	0.02	158	0.03	184	0.03
40	50	5	0.20	2	5	5	0.39	13	0.03	34	0.30	85	0.07
69	50	5	0.30	2	1	16	0.07	18	0.03	138	0.05	176	0.03
70	50	5	0.30	2	5	3	0.39	19	0.03	20	0.32	104	0.08
14	50	25	0.10	2	1	7	0.27	17	0.03	20	0.28	65	0.07
15	50	25	0.10	2	5	0	1.00	18	0.09	0	1.00	1	0.81
44	50	25	0.20	2	1	5	0.33	15	0.02	15	0.32	61	0.08
45	50	25	0.20	2	5	0	1.00	15	0.04	0	1.00	1	0.85
74	50	25	0.30	2	1	1	0.39	15	0.05	7	0.40	47	0.13
75	50	25	0.30	2	5	0	1.00	14	0.02	0	1.00	0	0.89
19	100	2	0.10	2	1	26	0.05	23	0.04	215	0.02	224	0.02
20	100	2	0.10	2	5	11	0.27	30	0.04	133	0.07	203	0.03
49	100	2	0.20	2	1	20	0.05	21	0.04	204	0.03	233	0.02
50	100	2	0.20	2	5	13	0.28	31	0.07	121	0.09	187	0.06
79	100	2	0.30	2	1	28	0.05	22	0.03	206	0.05	236	0.03
80	100	2	0.30	2	5	9	0.24	36	0.05	91	0.11	196	0.04
24	100	5	0.10	2	1	17	0.07	34	0.03	2	0.87	1	0.86
25	100	5	0.10	2	5	0	0.79	21	0.03	7	0.53	72	0.13
54	100	5	0.20	2	1	19	0.05	29	0.03	126	0.06	159	0.04
55	100	5	0.20	2	5	1	0.65	23	0.05	5	0.54	64	0.16
84	100	5	0.30	2	1	77	0.83	31	0.03	118	0.09	160	0.05
85	100	5	0.30	2	5	0	0.72	15	0.09	0	0.69	44	0.20
29	100	25	0.10	2	1	1	0.47	12	0.06	7	0.49	29	0.14
30	100	25	0.10	2	5	0	1.00	17	0.08	0	1.00	0	1.00
59	100	25	0.20	2	1	1	0.62	18	0.06	4	0.64	36	0.18
60	100	25	0.20	2	5	0	1.00	13	0.04	0	1.00	0	0.99
89	100	25	0.30	2	1	0	0.71	14	0.06	0	0.70	20	0.23
90	100	25	0.30	2	5	0	1.00	19	0.05	0	1.00	0	0.99

L2N, level-2 sample size; L1N, level −1 sample size; ICC, intraclass correlation; Items, number of biased items; NLD, count of negative log-likelihood difference test results; TP, true positive rate; FP, false positive rate.

(power) and false positive (Type I error) was partially supported. While the free vs. fixed distinction was unrelated to the true positive rate, the free vs. fixed distinction was related to the false positive rate. Hypothesis 2 proposed that the improved decision accuracy under the free baseline approach would be greater under conditions that should lead to greater power and lower Type I error, such as increased ICC, level-2 sample size, level-1 sample size, number of biased items and bias magnitude. Indeed, several moderation effects were observed.

Exploration of the interaction between free vs. fixed baseline approach and L1N revealed that the constrained approach led to increased false positive rates at increased L1N. This is a theoretically interpretable result. The increased L1N is expected to magnify the power to detect the misspecified larger model under the constrained baseline approach, a misspecification that contravenes the assumption of the log-likelihood difference test and that is not present under the free baseline approach. The interaction between the free vs. fixed approach and the number of biased items also indicated that as the number of biased items increased from 1 to 2 items, the false positive rate increased. Once again, this can be considered theoretically consistent, because under the constrained baseline approach

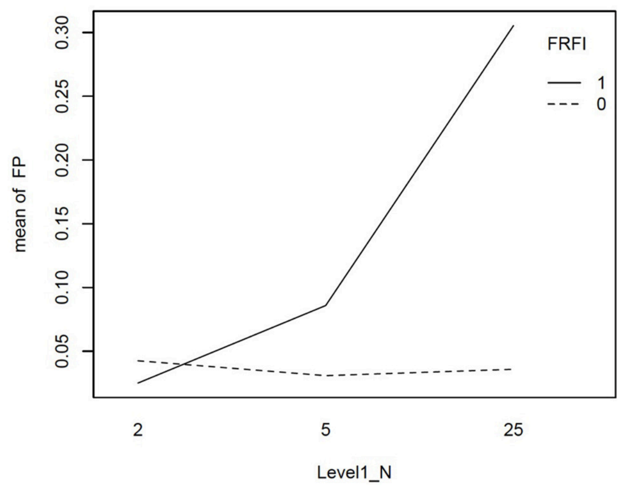


FIGURE 2 | Interaction plot for the effect of level-1 sample size and the free vs. fixed baseline choice on the false positive rate when the referent indicator is unbiased. FRFI, free vs. fixed, where 0 is free and 1 is fixed; FP, false positive rate; L1N, level 1 sample size; where 2, 5, and 25 are the level 1 sample sizes studied.

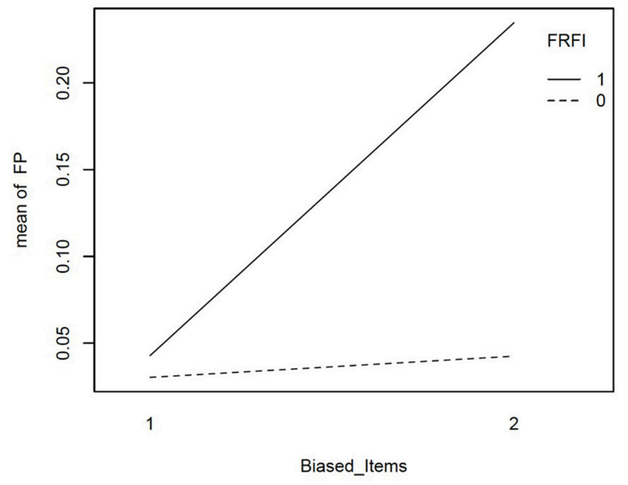


FIGURE 3 | Interaction plot for the effect of number of biased items and the free vs. fixed baseline choice on the false positive rate when the referent indicator is unbiased. FRFI, free vs. fixed, where 0 is free and 1 is fixed; FP, false positive rate; Biased items, number of biased items where 1 = 1 biased item and 2 = 2 biased items.

the inclusion of an additional misspecified item increases the degree of misspecification in the unrestricted model, and a correctly specified unrestricted model is required for the log-likelihood difference test. The final interaction between the free vs. fixed approach and the size of the bias indicated that as the bias increased so did the false positive rate when moving to the constrained baseline condition. Since increasing the size of the bias makes the misspecification more readily detectable, the violation of the assumption of the test procedure is once again more salient. Perhaps trumping all of these considerations,

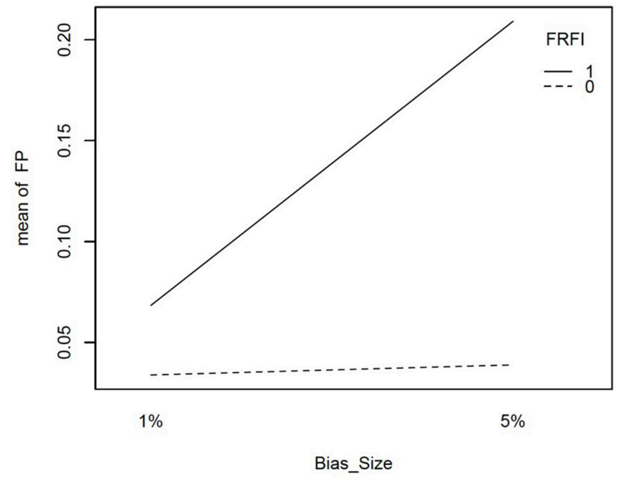


FIGURE 4 | Interaction plot for size if item bias and free vs. fixed baseline approach on the false positive rate when the referent indicator is unbiased. FRFI, free vs. fixed, where 0 is free and 1 is fixed; FP, false positive rate; Bias size, size of the bias, where 1% = 1% bias and 5% = 5% bias.

however, is the very high rate of negative log-likelihood difference test results under the constrained baseline set up. For all these reasons, the free baseline approach has more support when the anchor item is unbiased.

The measurement invariance literature has witnessed considerable research into the impact of violating the assumption of an unbiased indicator. This is because in practice, the assumption of an unbiased reference indicator might be easily violated. Therefore, we also examined what happens when the unbiased referent assumption is violated, since this assumption in practice can be difficult to check. Overall, when the referent indicator is biased, we saw that the false positive rate was unacceptably high regardless of whether a free or constrained baseline approach was used. The false positive rate for the free baseline approach was 0.25, and it was 0.29 for the fixed baseline approach. Moreover, the power was mediocre regardless of whether a free baseline approach or a fixed baseline approach was used. The power was 0.16 for the free baseline and 0.28 for the fixed baseline. These values make comparison of the advantages of one method over the other meaningless when the anchor item is biased. Instead, attention needs to be devoted to ensuring the referent item is unbiased. One approach that may be worthwhile considering is to first begin with the fully constrained model, and then examine modification indices to determine the item that is most likely to be unbiased. Once the unbiased item is identified, analyses can proceed according to the free baseline approach.

Researchers analyzing real data will need to make a series of decisions prior to the analysis and decisions during the analysis that will impact their ability to detect cluster bias. In terms of design considerations, this study reveals when the anchor item is unbiased, if the number of level-2 clusters is sufficiently large, increasing the level-1 sample size increases decision accuracy more than increasing level-2 sample size.

TABLE 4 | True positive and false positive rates for biased anchor with one additional biased item.

						Free baseline				Fixed baseline			
Cell	**L2N**	**L1N**	**ICC**	**Items**	**Size (%)**	**NLD**	**TP**	**NLD**	**FP**	**NLD**	**TP**	**NLD**	**FP**
91	50	2	0.10	1	1	9	0.03	8	0.04	206	0.03	217	0.01
92	50	2	0.10	1	5	6	0.07	9	0.10	160	0.07	203	0.03
103	50	2	0.20	1	1	19	0.05	20	0.03	202	0.00	233	0.00
104	50	2	0.20	1	5	7	0.10	7	0.15	164	0.07	215	0.04
115	50	2	0.30	1	1	18	0.06	24	0.04	201	0.03	228	0.03
116	50	2	0.30	1	5	17	0.13	11	0.15	139	0.07	166	0.06
93	50	5	0.10	1	1	20	0.03	21	0.04	175	0.03	195	0.03
94	50	5	0.10	1	5	17	0.03	6	0.37	50	0.26	106	0.06
105	50	5	0.20	1	1	19	0.03	18	0.03	151	0.05	158	0.02
106	50	5	0.20	1	5	13	0.12	13	0.24	32	0.28	116	0.11
117	50	5	0.30	1	1	25	0.03	17	0.05	147	0.04	183	0.03
118	50	5	0.30	1	5	5	0.20	14	0.15	14	0.37	93	0.11
95	50	25	0.10	1	1	12	0.17	8	0.09	15	0.31	43	0.12
96	50	25	0.10	1	5	61	0.62	7	0.20	0	1.00	0	0.81
107	50	25	0.20	1	1	1	0.28	7	0.07	13	0.35	47	0.10
108	50	25	0.20	1	5	0	0.97	11	0.09	0	1.00	0	0.88
119	50	25	0.30	1	1	3	0.33	14	0.06	5	0.38	41	0.11
120	50	25	0.30	1	5	0	1.00	13	0.07	0	1.00	0	0.89
97	100	2	0.10	1	1	20	0.04	25	0.05	239	0.01	217	0.01
98	100	2	0.10	1	5	11	0.10	11	0.19	137	0.07	227	0.02
109	100	2	0.20	1	1	24	0.04	18	0.07	243	0.03	216	0.03
110	100	2	0.20	1	5	12	0.16	8	0.22	115	0.10	168	0.03
121	100	2	0.30	1	1	34	0.04	30	0.04	196	0.04	229	0.03
122	100	2	0.30	1	5	15	0.18	12	0.21	78	0.12	162	0.06
99	100	5	0.10	1	1	28	0.03	24	0.04	150	0.04	184	0.03
100	100	5	0.10	1	5	17	0.03	2	0.56	14	0.47	51	0.14
111	100	5	0.20	1	1	24	0.02	20	0.04	129	0.06	140	0.03
112	100	5	0.20	1	5	8	0.26	13	0.27	7	0.56	49	0.17
123	100	5	0.30	1	1	20	0.06	26	0.05	94	0.08	141	0.04
124	100	5	0.30	1	5	2	0.43	11	0.14	2	0.65	54	0.19
101	100	25	0.10	1	1	3	0.41	12	0.10	4	0.52	40	0.18
102	100	25	0.10	1	5	48	0.76	10	0.21	0	1.00	0	0.99
113	100	25	0.20	1	1	2	0.55	12	0.08	1	0.61	26	0.17
114	100	25	0.20	1	5	0	1.00	4	0.10	0	1.00	0	0.99
125	100	25	0.30	1	1	1	0.68	18	0.06	1	0.73	29	0.18
126	100	25	0.30	1	5	0	1.00	16	0.07	0	1.00	0	0.99

L2N, level-2 sample size; L1N, level −1 sample size; ICC, intraclass correlation; Items, number of biased items; NLD, count of negative log-likelihood difference test results; TP, true positive rate; FP, false positive rate.

In addition, the ICC sizes studied had a negligible impact on decision accuracy. This is fortunate, sample sizes may be more under the researcher's control than ICCs. This conclusion, of course, is conditional on an item being identified as unbiased for identification.

The limitations of this study relate primarily to the inability to be exhaustive in the simulation conditions, for instance, with a wider range of L2N. Our results also only focus on continuous variable measurement models, and conclusions may not apply for ordered categorical items where the free baseline model used here may not converge (early experimentation has indicated that the level-2 residual variances need to be fixed at zero for models to converge). Moreover, following this study there are still important questions to investigate. There are numerous other constrained baseline approaches that might be considered, and this study does not speak to these methods. For example, alternative constrained baseline methods may well perform better than the constrained baseline approach used here. These could include iterative freeing of residual variances based on modification indices, simultaneous freeing of all residual variances followed by determining significance with standard errors, or freeing residual variances one by one, and leaving

TABLE 5 | True positive and false positive rates for biased anchor with two additional biased item.

						Free baseline				Fixed baseline			
Cell	**L2N**	**L1N**	**ICC**	**Items**	**Size (%)**	**NLD**	**TP**	**NLD**	**FP**	**NLD**	**TP**	**NLD**	**FP**
127	50	2	0.10	2	1	19	0.03	13	0.06	325	0.29	324	0.30
128	50	2	0.10	2	5	14	0.03	9	0.16	181	0.01	156	0.04
139	50	2	0.20	2	1	15	0.04	18	0.06	236	0.03	203	0.03
140	50	2	0.20	2	5	16	0.05	9	0.14	216	0.03	152	0.05
151	50	2	0.30	2	1	21	0.04	11	0.04	231	0.02	204	0.01
152	50	2	0.30	2	5	137	0.28	134	0.31	197	0.03	127	0.07
129	50	5	0.10	2	1	14	0.03	16	0.04	193	0.03	189	0.03
130	50	5	0.10	2	5	9	0.02	1	0.55	121	0.07	27	0.33
141	50	5	0.20	2	1	19	0.03	18	0.05	164	0.02	159	0.04
142	50	5	0.20	2	5	27	0.02	2	0.44	102	0.09	18	0.35
153	50	5	0.30	2	1	30	0.03	13	0.04	159	0.04	123	0.03
154	50	5	0.30	2	5	23	0.01	1	0.42	112	0.06	29	0.37
131	50	25	0.10	2	1	17	0.04	3	0.30	64	0.09	18	0.26
132	50	25	0.10	2	5	16	0.02	0	1.00	0	0.79	0	1.00
143	50	25	0.20	2	1	19	0.04	3	0.32	58	0.09	14	0.31
144	50	25	0.20	2	5	5	0.04	0	1.00	0	0.79	0	1.00
155	50	25	0.30	2	1	15	0.04	2	0.44	43	0.13	8	0.44
156	50	25	0.30	2	5	7	0.02	0	1.00	0	0.85	0	1.00
133	100	2	0.10	2	1	25	0.05	23	0.04	230	0.02	226	0.02
134	100	2	0.10	2	5	16	0.06	5	0.29	230	0.02	146	0.07
145	100	2	0.20	2	1	29	0.04	24	0.03	235	0.03	218	0.02
146	100	2	0.20	2	5	22	0.09	12	0.29	191	0.03	91	0.11
157	100	2	0.30	2	1	39	0.03	23	0.04	243	0.04	222	0.02
158	100	2	0.30	2	5	31	0.11	7	0.25	197	0.04	87	0.12
135	100	5	0.10	2	1	28	0.02	16	0.05	168	0.03	150	0.04
136	100	5	0.10	2	5	21	0.03	0	0.82	70	0.15	8	0.60
147	100	5	0.20	2	1	26	0.02	14	0.05	175	0.02	112	0.06
148	100	5	0.20	2	5	21	0.04	0	0.73	63	0.16	3	0.63
159	100	5	0.30	2	1	20	0.04	18	0.07	157	0.04	106	0.05
160	100	5	0.30	2	5	11	0.04	0	0.79	52	0.15	0	0.70
137	100	25	0.10	2	1	16	0.04	0	0.55	43	0.15	4	0.53
138	100	25	0.10	2	5	12	0.03	0	1.00	0	0.98	0	1.00
149	100	25	0.20	2	1	19	0.05	1	0.63	27	0.19	2	0.63
150	100	25	0.20	2	5	15	0.03	15	0.03	0	0.98	0	1.00
161	100	25	0.30	2	1	16	0.04	0	0.76	28	0.18	2	0.76
162	100	25	0.30	2	5	18	0.02	0	1.00	0	1.00	0	1.00

L2N, level-2 sample size; L1N, level −1 sample size; ICC, intraclass correlation; Items, number of biased items; NLD, count of negative log-likelihood difference test results; TP, true positive rate; FP, false positive rate.

the residual variance free for the item with the largest chi-square difference. Another avenue for research could be to examine whether the power of the cluster bias test increases when the likelihood ratio test distribution is adjusted to account for the level-2 residual variance test examining the boundary of the admissible parameter space (Stoel et al., 2006).

In summary, this study supports the free baseline approach when model assumptions are met. These might include situations where well developed psychometric instruments have been independently used in many different countries, and we know for instance, that similar items have corresponding high factor loadings in the different countries from independent research. In these instances, the lower false positive rate for the free baseline approach should lead to its adoption over the constrained baseline approach.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and approved it for publication.

REFERENCES

Asparouhov, T., and Muthén, B. (2014). Multiple-group factor analysis alignment. *Struct. Equ. Model. Multidiscip. J.* 21, 495–508. doi: 10.1080/10705511.2014.919210

Barendse, M., Oort, F., Werner, C., Ligtvoet, R., and Schermelleh-Engel, K. (2012). Measurement bias detection through factor analysis. *Struct. Equ. Model. Multidiscip. J.* 19, 561–579. doi: 10.1080/10705511.2012.713261

Barendse, M. T., Oort, F. J., and Garst, G. J. (2010). Using restricted factor analysis with latent moderated structures to detect uniform and nonuniform measurement bias; a simulation study. *AStA Adv. Stat. Anal.* 94, 117–127. doi: 10.1007/s10182-010-0126-1

Byrne, B., M., Shavelson, R. J., and Muthen, B. (1989). Testing for the equivalence of factor covariance and mean structures: the issue of partial measurement invariance. *Psychol. Bull.* 105, 456–466. doi: 10.1037/0033-2909.105.3.456

Cham, H., West, S. G., Ma, Y., and Aiken, L. S. (2012). Estimating latent variable interactions with nonnormal observed data: a comparison of four approaches. *Multivariate Behav. Res.* 47, 840–876. doi: 10.1080/00273171.2012.732901

Chan, D. (1998). The conceptualization and analysis of change over time: an integrative approach incorporating longitudinal mean and covariance structures analysis (LMACS) and multiple indicator latent growth modeling (MLGM). *Organ. Res. Methods* 1, 421–483. doi: 10.1177/109442819814004

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *J. Pers. Soc. Psychol.* 95, 1005–1018. doi: 10.1037/a0013193

Cheng, C., Cheung, M. W. L., and Montasem, A. (2014). Explaining differences in subjective well-being across 33 nations using multilevel models: universal personality, cultural relativity, and national income. *J. Pers.* 84, 46–58. doi: 10.1111/jopy.12136

Cheung, G. W., and Rensvold, R. B. (1999). Testing factorial invariance across groups: a reconceptualization and proposed new method. *J. Manage* 25, 1–27. doi: 10.1177/014920639902500101

Cheung, M. W. L., Leung, K., and Au, K. (2006). Evaluating multilevel models in cross-cultural research: an illustration with social axioms. *J. Cross Cult. Psychol.* 37, 522–541. doi: 10.1177/0022022106290476

Chun, S., Stark, S., Kim, E. S., and Chernyshenko, O. S. (2016). MIMIC Methods for Detecting DIF Among Multiple Groups: exploring a New Sequential-Free Baseline Procedure. *Appl. Psychol. Meas.* 40, 486–499. doi: 10.1177/0146621616659738

Cieciuch, J., Davidov, E., Algesheimer, R., and Schmidt, P. (2017). Testing for approximate measurement invariance of human values in the European Social Survey. *Sociol. Methods Res.* doi: 10.1177/0049124117701478. [Epub ahead of print].

Cieciuch, J., Shalom, S., Davidov, E., Schmidt, P., and Algesheimer, R. (2014). Comparing results of an exact vs. an approximate (Bayesian) measurement invariance test: a cross-country illustration with a scale to measure 19 human values. *Front. Psychol.* 5:982. doi: 10.3389/fpsyg.2014.00982

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences, 2nd Edn*. Hillsdale, NJ: Routledge Academic.

Drasgow, F. (1982). Biased test items and differential validity. *Psychol. Bull.* 92, 526–531. doi: 10.1037/0033-2909.92.2.526

Drasgow, F. (1984). Scrutinizing psychological tests: measurement equivalence and equivalent relations with external variables are the central issues. *Psychol. Bull.* 95, 134–135. doi: 10.1037/0033-2909.95.1.134

Guenole, N. (2016). The importance of isomorphism for conclusions about homology: a Bayesian multilevel structural equation modeling approach with ordinal indicators. *Front. Psychol.* 7:289. doi: 10.3389/fpsyg.2016.00289

Guenole, N., and Brown, A. (2014). The consequences of ignoring measurement invariance for path coefficients in structural equation models. *Front. Psychol.* 5:980. doi: 10.3389/fpsyg.2014.00980

Jak, S., and Oort, F. J. (2015). On the power of the test for cluster bias. *Br. J. Math. Stat. Psychol.* 68, 434–455. doi: 10.1111/bmsp.12053

Jak, S., Oort, F. J., and Dolan, C. V. (2013). A test for cluster bias: detecting violations of measurement invariance across clusters in multilevel data. *Struct. Equ. Modeling* 20, 265–282. doi: 10.1080/10705511.2013.769392

Jak, S., Oort, F. J., and Dolan, C. V. (2014). Measurement bias in multilevel data. *Struct. Equ. Modeling* 21, 31–39. doi: 10.1080/10705511.2014.856694

Jang, S., Kim, E. S., Cao, C., Allen, T. D., Cooper, C. L., Lapierre, L. M., et al. (2017). Measurement invariance of the satisfaction with life scale across 26 countries. *J. Cross Cult. Psychol.* 48, 560-576. doi: 10.1177/0022022117697844

Joreskog, K. G., and Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *J. Am. Stat. Assoc.* 70, 631–639.

Kim, E. S., Kwok, O. M., and Yoon, M. (2012a). Testing factorial invariance in multilevel data: a Monte Carlo study. *Struct. Equ. Model. Multidiscip. J.* 19, 250–267. doi: 10.1080/10705511.2012.659623

Kim, E. S., Yoon, M., and Lee, T. (2012b). Testing measurement invariance using MIMIC: likelihood ratio test with a critical value adjustment. *Educ. Psychol. Meas.* 72, 469–492. doi: 10.1177/0013164411427395

Maas, C. J. M., and Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology* 1, 86–92. doi: 10.1027/1614-1881.1.3.86

Mellenbergh, G. J. (1989). Item bias and item response theory. *Int. J. Educ. Res.* 13, 127–143. doi: 10.1016/0883-0355(89)90002-5

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika* 58, 525–543. doi: 10.1007/BF02294825

Millsap, R. E. (2012). *Statistical Approaches to Measurement Invariance*. London: Routledge.

Muthén, B., Khoo, S. T., and Gustafsson, J. E. (1997). *Multilevel Latent Variable Modeling in Multiple Populations*. Unpublished Technical Report. Available Online at: http://www.statmodel.com

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *J. Edu. Meas.* 28, 338–354. doi: 10.1111/j.1745-3984.1991.tb00363.x

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociol. Methods Res.* 22, 376–398. doi: 10.1177/0049124194022003006

Muthén, L. K., and Muthén, B. O. (1998-2017). *Mplus User's Guide, 8th Edn*. Los Angeles, CA: Muthén & Muthén.

Navas-Ara, M. J., and Gómez-Benito, J. (2002). Effects of ability scale purification on identification of DIF. *Eur. J. Psychol. Asses.* 18, 9–15. doi: 10.1027//1015-5759.18.1.9

Oort, F. J. (1998). Simulation study of item bias detection with restricted factor analysis. *Struct. Equ. Modeling Multidiscip. J.* 5, 107–124. doi: 10.1080/10705519809540095

Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., and Chen, F. (2001). Monte Carlo experiments: design and implementation. *Struct. Equ. Modeling* 8, 287–312. doi: 10.1207/S15328007SEM0802_7

Rabe-Hesketh, S., Skrondal, A., and Pickles, A. (2004). Generalized multilevel structural equation modelling. *Psychometrika* 69, 167–190. doi: 10.1007/BF02295939

Ryu, E. (2014). Factorial invariance in multilevel confirmatory factor analysis. *Br. J. Math. Stat. Psychol.* 67, 172–194. doi: 10.1111/bmsp.12014

Satorra, A., and Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika* 66, 507–514. doi: 10.1007/BF02296192

Sorbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *Br. J. Math. Stat. Psychol.* 27, 229–239. doi: 10.1111/j.2044-8317.1974.tb00543.x

Stark, S., Chernyshenko, O. S., and Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: toward a unified strategy. *J. Appl. Psychol.* 91, 1292–1306. doi: 10.1037/0021-9010.91.6.1292

Stoel, R. D., Garre, F. G., Dolan, C., and van den Wittenboer, G. (2006). On the likelihood ratio test in structural equation modeling when parameters are subject to boundary constraints. *Psychol. Methods* 11, 439–455. doi: 10.1037/1082-989X.11.4.439

Tay, L., Woo, S. E., and Vermunt, J. K. (2014). A conceptual framework of cross-level Isomorphism: psychometric validation of multilevel constructs. *Organ. Res. Methods* 17, 77–106. doi: 10.1177/1094428113517008

Vandenberg, R. J., and Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. *Organ. Res. Methods* 3, 4–70. doi: 10.1177/109442810031002

van de Schoot, A. G. J., Schmidt, P., and De Beuckelaer, A. (2015). *Measurement Invariance.* Lausanne: Front. Media.

Wang, W., Tay, L., and Drasgow, F. (2013). Detecting differential item functioning of polytomous items for an ideal point response process. *Appl. Psychol. Meas.* 37, 316–335. doi: 10.1177/0146621613476156

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

METHODS
published: 17 August 2018
doi: 10.3389/fpsyg.2018.01387

Computing Multivariate Effect Sizes and Their Sampling Covariance Matrices With Structural Equation Modeling: Theory, Examples, and Computer Simulations

*Mike W.-L. Cheung**

Department of Psychology, National University of Singapore, Singapore, Singapore

In the social and behavioral sciences, it is recommended that effect sizes and their sampling variances be reported. Formulas for common effect sizes such as standardized and raw mean differences, correlation coefficients, and odds ratios are well known and have been well studied. However, the statistical properties of multivariate effect sizes have received less attention in the literature. This study shows how structural equation modeling (SEM) can be used to compute multivariate effect sizes and their sampling covariance matrices. We focus on the standardized mean difference (multiple-treatment and multiple-endpoint studies) with or without the assumption of the homogeneity of variances (or covariance matrices) in this study. Empirical examples were used to illustrate the procedures in R. Two computer simulation studies were used to evaluate the empirical performance of the SEM approach. The findings suggest that in multiple-treatment and multiple-endpoint studies, when the assumption of the homogeneity of variances (or covariance matrices) is questionable, it is preferable not to impose this assumption when estimating the effect sizes. Implications and further directions are discussed.

Keywords: effect size, multivariate effect size, sampling covariance matrix, meta-analysis, structural equation model

OPEN ACCESS

Edited by:
Prathiba Natesan,
University of North Texas,
United States

Reviewed by:
Lietta Marie Scott,
Arizona Department of Education,
United States
Ehri Ryu,
Boston College, United States
Joshua N. Pritikin,
Virginia Commonwealth University,
United States

****Correspondence:***
Mike W.-L. Cheung
mikewlcheung@nus.edu.sg

Specialty section:
This article was submitted to
Quantitative Psychology and
Measurement,
a section of the journal
Frontiers in Psychology

Received: *31 July 2017*
Accepted: *17 July 2018*
Published: *17 August 2018*

Citation:
Cheung MW-L (2018) Computing Multivariate Effect Sizes and Their Sampling Covariance Matrices With Structural Equation Modeling: Theory, Examples, and Computer Simulations.
Front. Psychol. 9:1387.
doi: 10.3389/fpsyg.2018.01387

In the social and behavioral sciences, it is recommended that effect sizes and their sampling variances be reported (e.g., Cohen, 1994; Wilkinson and Task Force on Statistical Inference, 1999; Cumming, 2014). When there are a sufficient number of studies, the meta-analysis is the standard method used to synthesize the research findings. The results of the meta-analysis may inform us what the average effect is and how the effect sizes vary across the studies.

There are two key ingredients for a meta-analysis. The first one is the effect size that quantifies the strength of the effect in the studies. Effect sizes can be either unstandardized or standardized (e.g., Kelley and Preacher, 2012). Unstandardized effect sizes are used when the effect sizes are comparable across studies, e.g., blood pressure or physical measures (Bond et al., 2003). When the scales of the measures are unclear or non-comparable across studies, standardized effect sizes are preferred (e.g., Hunter and Hamilton, 2002).

Besides the effect sizes, we also need the standard error (*SE*) of the effect sizes to quantify the precision of the estimated effect sizes. Formulas for common effect sizes such as the standardized and raw mean differences, correlation coefficients, and odds ratios are well known and have been well studied (Borenstein et al., 2009; Card, 2012; Cheung, 2015a; Schmidt and Hunter, 2015).

In applied research, however, more than one effect size may be involved. For example, there may be more than one treatment group compared to a control group. The use of multiple treatment groups allows researchers to address the phenomenon under different levels of manipulation. By using the same control group in the comparisons, researchers minimize the cost of collecting multiple control groups (Kim and Becker, 2010). Another example is when there is more than one outcome variable in the control and treatment groups. The use of multiple outcomes permits researchers to study different related outcomes under the same manipulations (Thompson and Becker, 2014). Studies that measure these two types of effect sizes are known as multiple-treatment and multiple-endpoint studies.

Since the effect sizes are not independent, researchers have to calculate the sampling covariances among the effect sizes. Gleser and Olkin (1994, 2009) have provided the most comprehensive treatment of this subject to date. They derived formulas to compute the effect sizes and their sampling variances and covariances. Once the effect sizes and their sampling covariance matrices are available, a multivariate meta-analysis (Nam et al., 2003; Jackson et al., 2011; Cheung, 2013) can be performed on all effect sizes.

Although Gleser and Olkin (1994, 2009) have provided standard formulas to compute the effect sizes and their sampling covariance matrices for multiple-treatment and multiple-endpoint studies, there are a few limitations in their approach. First, it is not easy for users, especially those without a strong statistical background, to comprehend the logic in calculating the variances and covariances. Second, these formulas rely on the assumption of the homogeneity of variances or covariance matrices. Although it is possible to drop these assumptions, the derivations are not apparent for most users. Most users would just adopt these assumptions without considering the alternatives. Third, it is difficult to extend their formulas to more complicated cases. One such example is the combination of multiple-treatment with multiple-endpoint studies in the same publication. Many researchers simplify the effect sizes to either the multiple-treatment study or the multiple-endpoint study, which is not ideal because of the loss of information.

Structural equation modeling (SEM) is a favorite tool to use in analyzing multivariate data. It has been used to calculate *SE*s and confidence intervals for various effect sizes and indices (Raykov, 2001; Cheung and Chan, 2004; Preacher, 2006). Recently, Cheung (2015a, Chapter 3) showed how common effect sizes, including those in multiple-treatment and multiple-endpoint studies, and their sampling variances and covariances, can be computed using the SEM framework.

The SEM approach provides a graphical model of means, standard deviations, and correlations. The effect sizes are defined as functions of these parameters. Readers can get a better understanding of what these effect sizes mean. Second, assumptions of the homogeneity of variances, covariances, or correlations can be imposed or relaxed by the use of equality constraints on the parameters. By using the delta method built into the SEM packages, appropriate sampling covariance matrices can be automatically derived. Third, it is feasible to extend the SEM approach to more complicated situations. For example, the SEM approach can be used to calculate the effect sizes and their sampling covariance matrix for a combination of multiple-treatment and multiple-endpoint studies[1] The key advantage of this is that researchers only need to focus on the conceptual "definition" of the effect sizes; the sampling covariance matrix of the effect sizes is numerically calculated by the SEM packages.

The rest of this article is structured as follows. The next section contains a brief introduction on how to compute the effect sizes and their sampling covariance matrices for the multiple-treatment and multiple-endpoint designs in SEM. Two empirical examples are used to illustrate how to conduct the analyses using the metaSEM package (Cheung, 2015b) implemented in the R statistical platform (R Development Core Team, 2018). Two computer simulations are then presented to evaluate the empirical performance of the SEM approach under several conditions. Based on the findings of the simulation, this paper concludes that it is preferable not to impose the assumption of the homogeneity of variances (or covariances) when calculating the effect sizes for multiple treatment and multiple-endpoint studies when this assumption is questionable. Finally, further directions for further research are discussed.

A SEM APPROACH TO ESTIMATING EFFECT SIZE

Cheung (2015a, Chapter 3) presents a SEM approach to estimating various effect sizes, including those in multiple-treatment and multiple-endpoint studies. There are three steps in the analysis. In the first step, a structural equation model with means, standard deviations, and correlations is proposed to

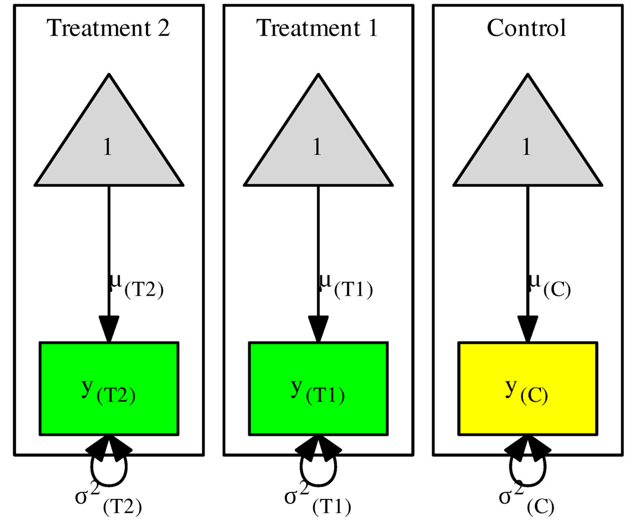


FIGURE 1 | The structural equation model for the multiple-treatment studies.

[1] https://stats.stackexchange.com/questions/108248/calculating-effect-sizes-and-standard-errors-for-the-difference-between-two-stan/130512.

fit the data. When the data are from independent groups (e.g., control vs. intervention groups in calculating the standardized or raw mean differences) a multiple-group structural equation model is used. Second, appropriate equality constraints on the homogeneity of covariance (or correlation) matrices are imposed. If there are reasons to believe that the assumption of the homogeneity of covariance (or correlation) matrices is not appropriate, researchers may test the hypothesis statistically. They may then choose to drop these assumptions when calculating the effect sizes.

Finally, the effect sizes are defined as functions of the means and standard deviations (*SD*s). The effect sizes with their sampling covariance matrices are estimated by the SEM packages using maximum likelihood (ML) estimation. This approach releases users from the need to manually derive the sampling covariance matrix, a process that is prone to human error. Let us consider examples of multiple-treatment and multiple-endpoint studies.

Multiple-Treatment Studies

Suppose that we measure the mathematics score in a control group and two treatment groups ($y_{(C)}$,$y_{(T1)}$, and $y_{(T2)}$). **Figure 1** shows a structural equation model with one control and two treatment groups. For ease of discussion, we use the population parameters in the figures. It is understood that sample estimates are employed in the analyses. The rectangles and the triangles represent the observed variables and columns of ones, respectively. The arrows from the triangles to the observed variables represent the means of the variables in the control $\mu_{(C)}$, treatment 1 $\mu_{(T1)}$, and treatment 2 $\mu_{(T2)}$, respectively. The variances of the variables in the control and treatments 1 and 2 are represented by $\sigma^2_{(C)}$, $\sigma^2_{(T1)}$, and $\sigma^2_{(T2)}$, respectively.

When no constraint is imposed, the above means and variances are the same as those of the sample statistics. Under the assumption of the homogeneity of variances, we may impose the constraint $H_0 : \sigma^2_{Common} = \sigma^2_{(C)} = \sigma^2_{(T1)} = \sigma^2_{(T2)}$. This null hypothesis is tested by comparing the likelihood ratio (*LR*) statistics of the models with and without the constraint. If the null hypothesis is correct, the difference between the *LR* statistics follows a chi-square distribution with 2 degrees of freedom (*df*s). We may now define the standardized mean differences (SMDs) between the treatment groups and the control by using the common *SD* σ_{Common} as the denominator:

$$SMD_{\mathrm{MTS1}} = \frac{\mu_{(T1)} - \mu_{(C)}}{\sigma_{\mathrm{Common}}} \text{ and } SMD_{MTS2} = \frac{\mu_{(T2)} - \mu_{(C)}}{\sigma_{\mathrm{Common}}}. \tag{1}$$

One unit of SMD indicates that the mean of the treatment group is one common *SD* above that of the control group. Since SMD_{MTS1} and SMD_{MTS2} share the same parameters $\mu_{(C)}$ and σ_{Common}, they are correlated. Instead of using the analytic solutions provided by Gleser and Olkin (1994, 2009), we may estimate the sampling variances and the covariance by the numerical approach in SEM.

When the assumption of the homogeneity of variances is questionable, it may not be appropriate to use σ_{Common} in the denominator. This is because σ_{Common} is not estimating any of the population *SD*s. A better alternative is to use the control group $\sigma_{(C)}$ as the standardizer in calculating the effect sizes

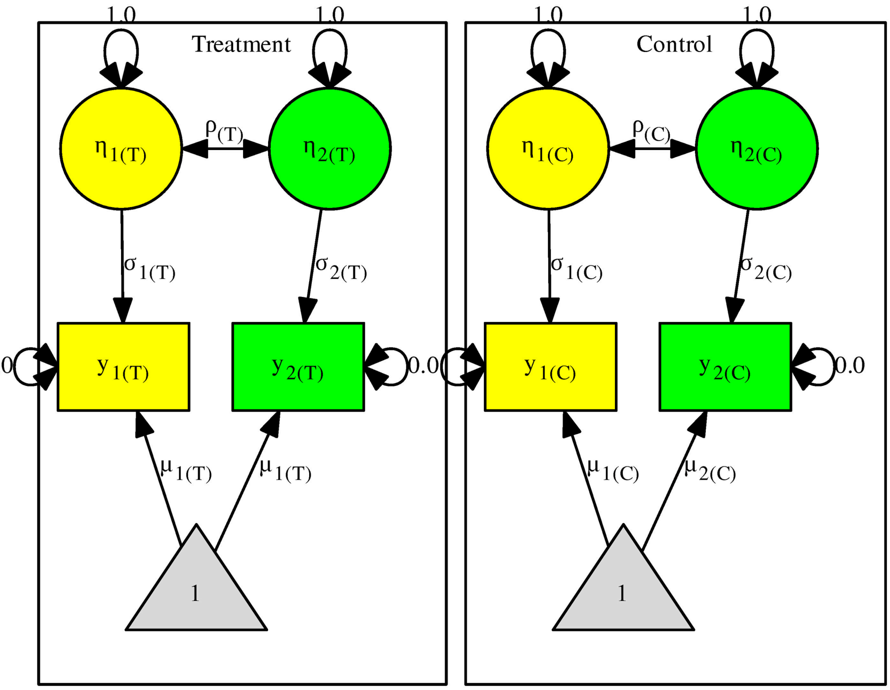


FIGURE 2 | The structural equation model for the multiple-endpoint studies.

(Glass et al., 1981). The standardized mean differences of the treatment groups against the control group are now described as:

$$SMD_{\text{MTS1}} = \frac{\mu_{(T1)} - \mu_{(C)}}{\sigma_{(C)}} \; \textit{and} \; SMD_{\text{MTS2}} = \frac{\mu_{(T2)} - \mu_{(C)}}{\sigma_{(C)}}, \tag{2}$$

which does not rely on the assumption of the homogeneity of variances. Now, one unit of SMD indicates that the mean of the treatment group is one *SD* of the control group above that of the control group.

Multiple-Endpoint Studies

Now suppose that there are two effect sizes on the mathematics and language scores y_1 and y_2. **Figure 2** shows the model with two independent groups (the control and treatment groups). We use η_1 and η_2, with their variances fixed at one, to represent the standardized scores of y_1 and y_2. σ_1 and σ_2 now represent the *SD*s of y_1 and y_2. The same model representation is often used to standardize the variables in SEM (e.g., Cheung and Chan, 2004, 2005; Cheung, 2015a).

We may assume that the correlations are homogeneous by imposing the constraint $H_0: \rho_{\text{Common}} = \rho_{(C)} = \rho_{(T)}$. An *LR* test can be used to test this null hypothesis by comparing the models with and without this constraint. Under the null hypothesis, the test statistic has a chi-square distribution with 1 *df*. If we further assume that the covariance matrices are homogeneous, we may impose the constraints of $H_0: \rho_{\text{Common}} = \rho_{(C)} = \rho_{(T)}$, $H_0: \sigma_{1\text{Common}} = \sigma_{1(C)} = \sigma_{1(T)}$, and $H_0: \sigma_{2\text{Common}} = \sigma_{2(C)} = \sigma_{2(T)}$. Under the null hypothesis, the test statistic on comparing the models with and without the constraints follows a chi-square distribution with 3 *df*s. We may drop all of the constraints if these assumptions are questionable.

Regardless of whether we have imposed the above constraints, the effect sizes for the multiple-endpoint study are defined as:

$$SMD_{\text{MES1}} = \frac{\mu_{1(T)} - \mu_{1(C)}}{\sigma_1} \; \textit{and} \; SMD_{\text{MES2}} = \frac{\mu_{2(T)} - \mu_{2(C)}}{\sigma_2}, \tag{3}$$

where σ_1 and σ_2 are the standard deviations for y_1 and y_2. We do not put the subscript in the formulas because what σ_1 and σ_2 actually are depends on whether constraints have been imposed on them. If we impose the equality constraints on the *SD*s, $\sigma_{1\text{Common}}$ and $\sigma_{2\text{Common}}$ are used as the standardizers in Equation (3). If we do not assume that

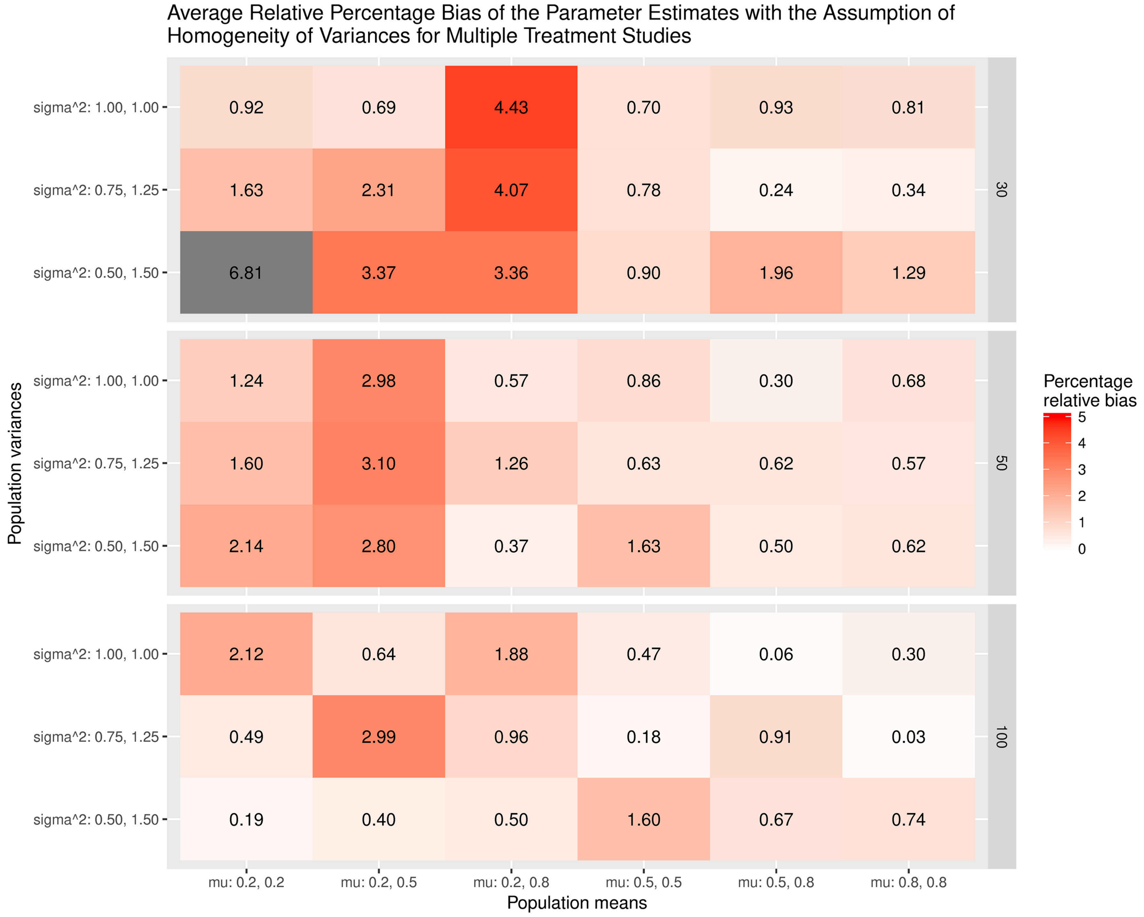


FIGURE 3 | Relative bias of the average of the parameter estimates for the multiple treatment studies with the assumption of homogeneity of variances.

the covariance matrices are homogeneous, the *SD*s in the control groups ($\sigma_{1(C)}$ and $\sigma_{2(C)}$) are used as the standardizers. Once we have defined the appropriate effect sizes, the sampling covariance matrix between SMD_{MES1} and SMD_{MES2} can be obtained from the SEM packages with numerical methods.

Illustrations With R

Gleser and Olkin (1994) presented some sample data on the multiple-treatment and multiple-endpoint studies. These datasets are stored in the metaSEM package (Cheung, 2015b). The metaSEM package also provides smdMTS() and smdMES() to calculate the effect sizes for a multiple-treatment study and a multiple-endpoint study with or without the assumptions of homogeneity. **Supplementary Materials 5** shows the sample R code. Readers may refer to the package manual for details.

Table 22.2 in Gleser and Olkin (1994) displays simulated data from six studies on five modes of exercise with a control group of no regular exercise. The dependent variable is systolic blood pressure. Therefore, a negative effect size between the treatment and control groups suggests that those in the treatment groups are in better health than those in the control group. As an illustration, we show the calculations from the first study, which includes three treatment groups and one control group. When we assume that the variances are homogeneous, the SMD_{MTS} of the three treatment groups compared to the control group are -1.17, -1.90, and -2.00, respectively. The sampling covariance matrix is $\begin{pmatrix} 0.09 & & \\ 0.05 & 0.10 & \\ 0.05 & 0.06 & 0.10 \end{pmatrix}$. If we do not assume that the variances are homogeneous and use the *SD* of the control group as the standardizer, the SMD_{MTS} are -0.79, -1.29, and -1.36, respectively. The sampling covariance matrix is $\begin{pmatrix} 0.06 & & \\ 0.06 & 0.09 & \\ 0.06 & 0.07 & 0.08 \end{pmatrix}$. In this example, the effect sizes that were calculated with the assumption that the variances are homogeneous and are about 50% larger than those that were calculated without this assumption. When testing the assumption that the variances are homogeneous, the statistic is $\chi^2_{(3)} = 21.30$, $p < 0.001$, which suggests that this assumption is not tenable. It is questionable whether the use of effect sizes with the assumption of the homogeneity of variances is appropriate in this example.

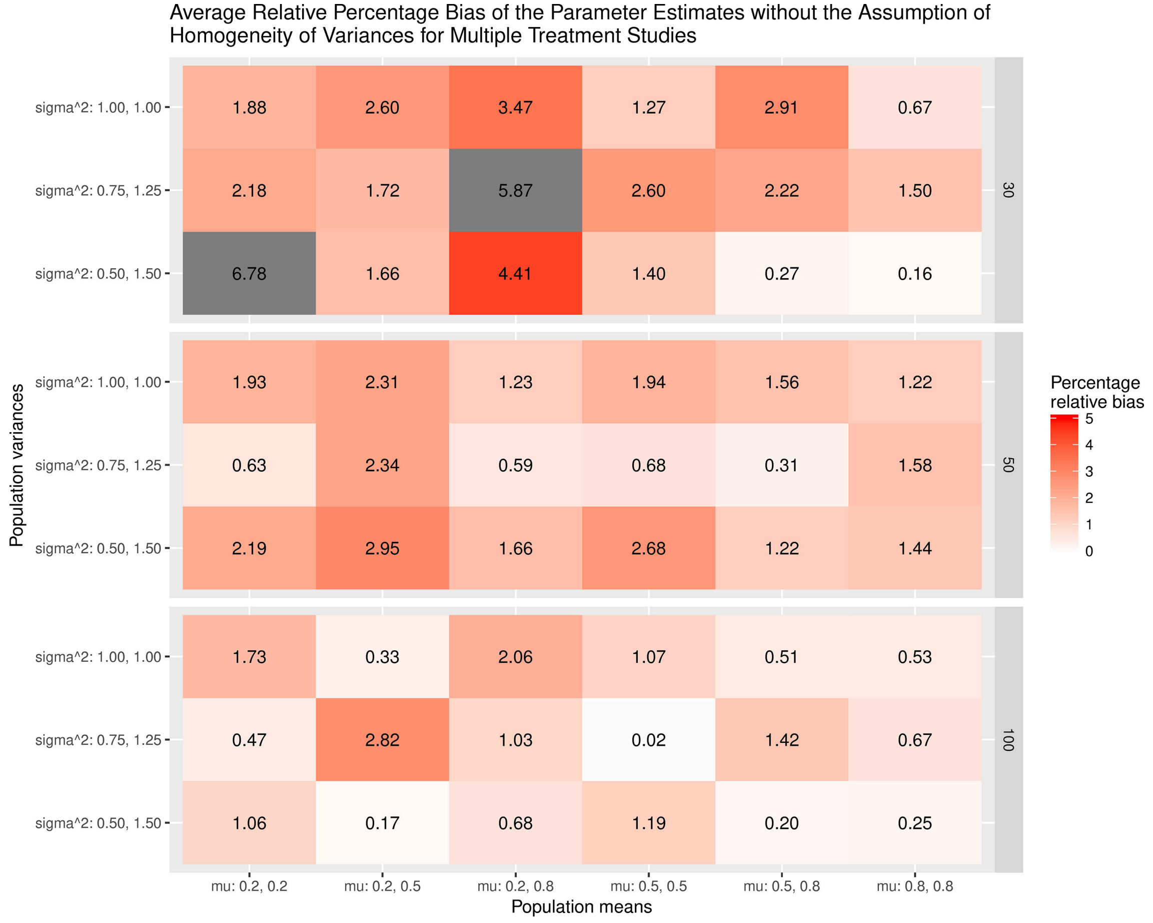


FIGURE 4 | Relative bias of the average of parameter estimates for the multiple treatment studies without the assumption of homogeneity of variances.

Table 22.4 in Gleser and Olkin (1994) shows seven published studies on the SAT-Math and SAT-Verbal scores of groups that had been coached on the tests compared to the scores of uncoached control groups. A positive effect size means that the coached groups performed better than the uncoached groups. As an illustration, we select the first study for demonstration. The SMD_{MES} on Math and Verbal are 1.19 and 0.61 with $V_{\text{MES}} = \begin{pmatrix} 0.09 & \\ 0.05 & 0.08 \end{pmatrix}$. If we do not assume that the covariance matrices are homogeneous, the SMD_{MES} on Math and Verbal are 1.30 and 0.56 with $V_{\text{MES}} = \begin{pmatrix} 0.12 & \\ 0.05 & 0.06 \end{pmatrix}$. The test statistic on the homogeneity of covariance matrices is $\chi^2_{(3)} = 4.92$, $p = 0.18$, which is not statistically significant. It should be noted that the sample sizes in these studies are quite small (at 34 and 21).

The above illustrations show that the effect sizes with and without the assumption of homogeneity may be very different depending on whether the homogeneity assumption holds. It remains unclear how these effect size estimates would work empirically in simulated data. The following computer simulation clarifies the empirical performance of these estimators.

TWO SIMULATION STUDIES

Two computer simulation studies were conducted to evaluate the empirical performance of the SEM approach. All of the simulations were performed with the metaSEM package (Cheung, 2015b) in the R statistical platform (R Development Core Team, 2018).

Before moving on to details of the simulation studies, it is essential to clarify the meanings of "with and without the homogeneity of variances (or covariance matrices)" in the simulation studies. The data are generated from either equal or unequal population variances (see the conditions of the Population Variances). Regardless of whether or not the population variances are equal, two sets of effect sizes are calculated from the same set of data—one assumes the homogeneity of variances, and the other does not.

When the data are generated from populations with equal variances, the effect sizes both with and without the homogeneity assumption should be correct. By assuming that the variances are homogeneous, which is correct in the generated data, the sampling variances of the effect sizes with the homogeneity assumption are usually smaller than those effect sizes without

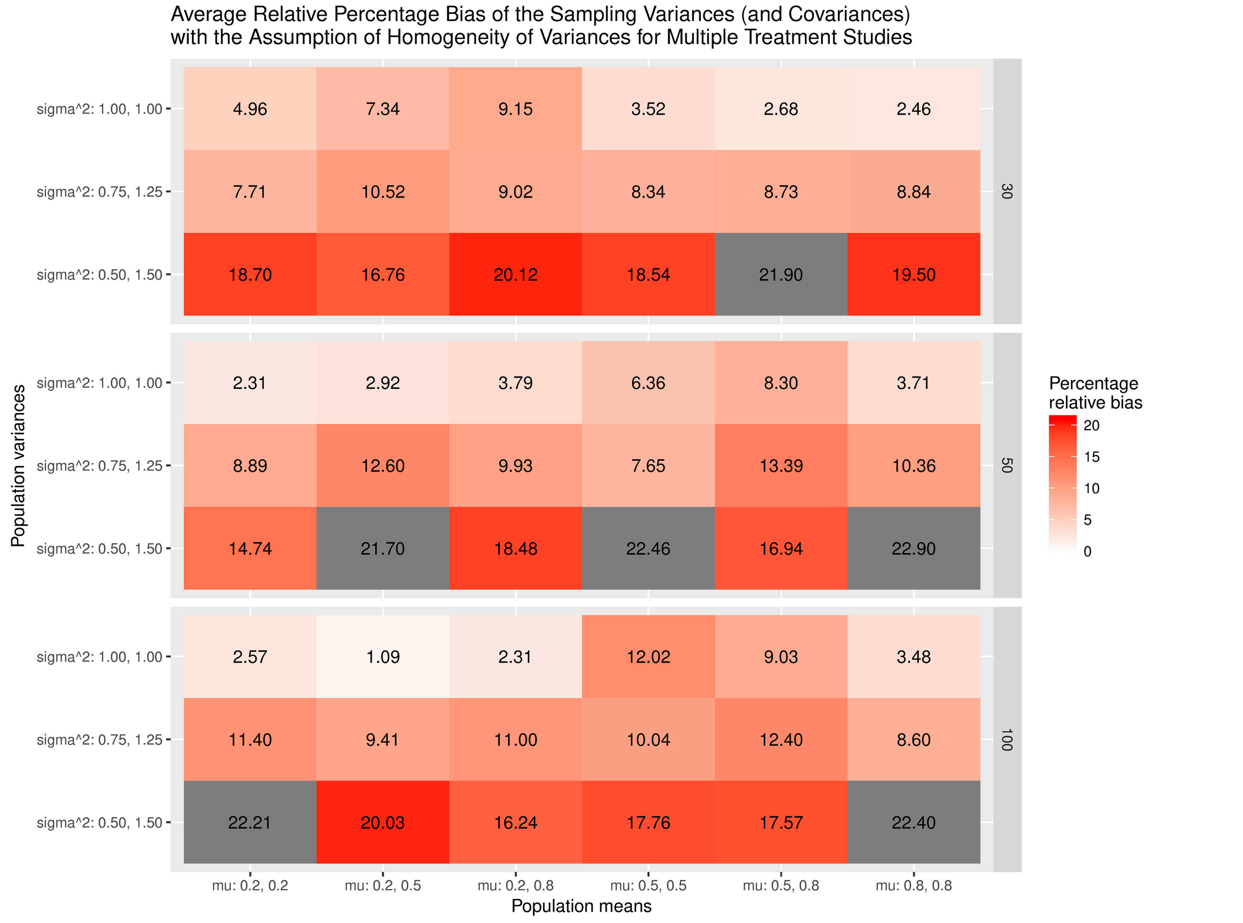


FIGURE 5 | Relative bias of the average of the sampling variances and covariance for the multiple treatment studies with the assumption of homogeneity of variances.

the homogeneity assumption. When the data are generated from unequal population variances, the effect sizes without the homogeneity assumption should still be correct. However, the effect sizes with the homogeneity assumption are likely to be biased because the model is misspecified. The present simulation studies evaluated the empirical performance of the computed effect sizes with and without the homogeneity assumption.

Study 1: Multiple-Treatment Studies

For the multiple-treatment studies, multivariate normal data were generated from the known data structures with or without the assumption of the homogeneity of variances.

Methods

In this simulation study, there was a control group with two treatment groups. Several factors were manipulated in the simulation study:

Population means

The population mean of the control group was fixed at 0 for reference. Six levels were used for the simulation study. The population means for the two treatment groups were (0.2, 0.2), (0.2, 0.5), (0.2, 0.8), (0.5, 0.5), (0.5, 0.8), and (0.8, 0.8).

Population variances

The population variance of the control group was fixed at 1 for reference. Three levels were selected for the simulation. The population variances for the two treatment groups were (1, 1), (0.75, 1.25), and (0.5, 1.5). When the population variance was (1, 1) in the two treatment groups, the homogeneity of variances was assumed. In the other levels, the population variances were heterogeneous. As the population variance of the control group was fixed at 1, the population effect size was calculated by the difference in means between the treatment groups and the control group divided by 1. Thus, the effect sizes were 0.2, 0.5, and 0.8, which represent the typical values observed in the social and behavioral sciences.

Sample sizes

The design was assumed to be balanced. Three levels of sample sizes were selected, namely, 30, 50, and 100. These levels should be representative of typical research settings.

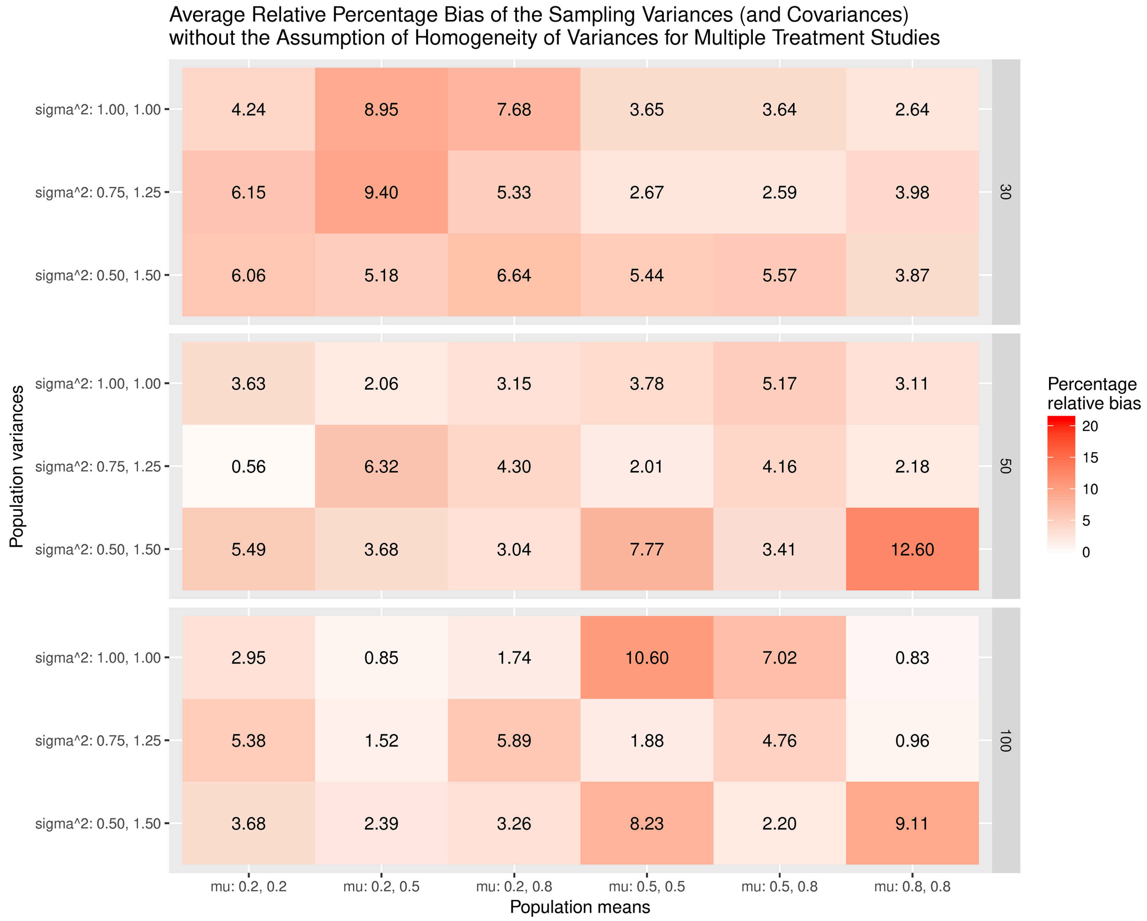


FIGURE 6 | Relative bias of the average of the sampling variances and covariance for the multiple treatment studies without the assumption of homogeneity of variances.

Thus, there were a total of 6 × 3 × 3 = 54 conditions. One thousand replications were repeated for each condition.

Assessment of the empirical performance

Since the population mean and variance of the control were set at 0 and 1, respectively, the population effect sizes were defined as the mean differences between the treatment 1 (or 2) to the control group. The relative percentage bias of each effect size was computed as

$$B(\hat{\theta}) = \frac{\bar{\hat{\theta}} - \theta}{\theta} \times 100\%, \tag{4}$$

where θ is the population effect size and $\bar{\hat{\theta}}$ is the mean of the estimates of the effect size $\hat{\theta}$ across the 1,000 replications. Proper estimation methods should have a relative bias of less than 5% (Hoogland and Boomsma, 1998). Since there were two effect sizes for two treatment groups, we reported the average of their absolute biases $B(\hat{\theta}) = \left(\left|B(\hat{\theta})_{\text{T1}}\right| + \left|B(\hat{\theta})_{\text{T2}}\right|\right)/2$, where $\left|B(\hat{\theta})_{\text{T1}}\right|$ and $\left|B(\hat{\theta})_{\text{T2}}\right|$ are the absolute biases for treatments 1 and 2, for ease of presentation.

When there is only one effect size, we may quantify the accuracy of its uncertainty by the use of the relative bias of the *SE*. Since there were two sampling variances and one sampling covariance, we used the sampling variances (SE^2) and covariance as the measure of uncertainty,

$$B\left(\overline{Var}(\hat{\theta})\right) = \frac{\overline{SE^2}(\hat{\theta}) - Var(\hat{\theta})}{Var(\hat{\theta})}, \tag{5}$$

where $Var(\hat{\theta})$ is the empirical variance (or covariance) of $\hat{\theta}$ and $\overline{SE^2}(\hat{\theta})$ is the mean of the SE^2 or sampling covariance across 1,000 replications. Since there were three biases for the two effect sizes and their covariance, we reported the average of their absolute biases.

$$\begin{aligned} B\left(\overline{Var}(\hat{\theta})\right) = {} & \left(\left|B\left(\overline{Var}(\hat{\theta}_1)\right)\right| + \left|B\left(\overline{Var}(\hat{\theta}_2)\right)\right|\right) \\ & + \left(\left|B\left(\overline{Cov}(\hat{\theta}_1, \hat{\theta}_2)\right)\right|\right)/3, \end{aligned} \tag{6}$$

where $\left|B\left(\overline{Var}(\hat{\theta}_1)\right)\right|$, $\left|B\left(\overline{Var}(\hat{\theta}_2)\right)\right|$ and $\left|B\left(\overline{Cov}(\hat{\theta}_1, \hat{\theta}_2)\right)\right|$ are the absolute biases for the outcomes 1 and 2 and their

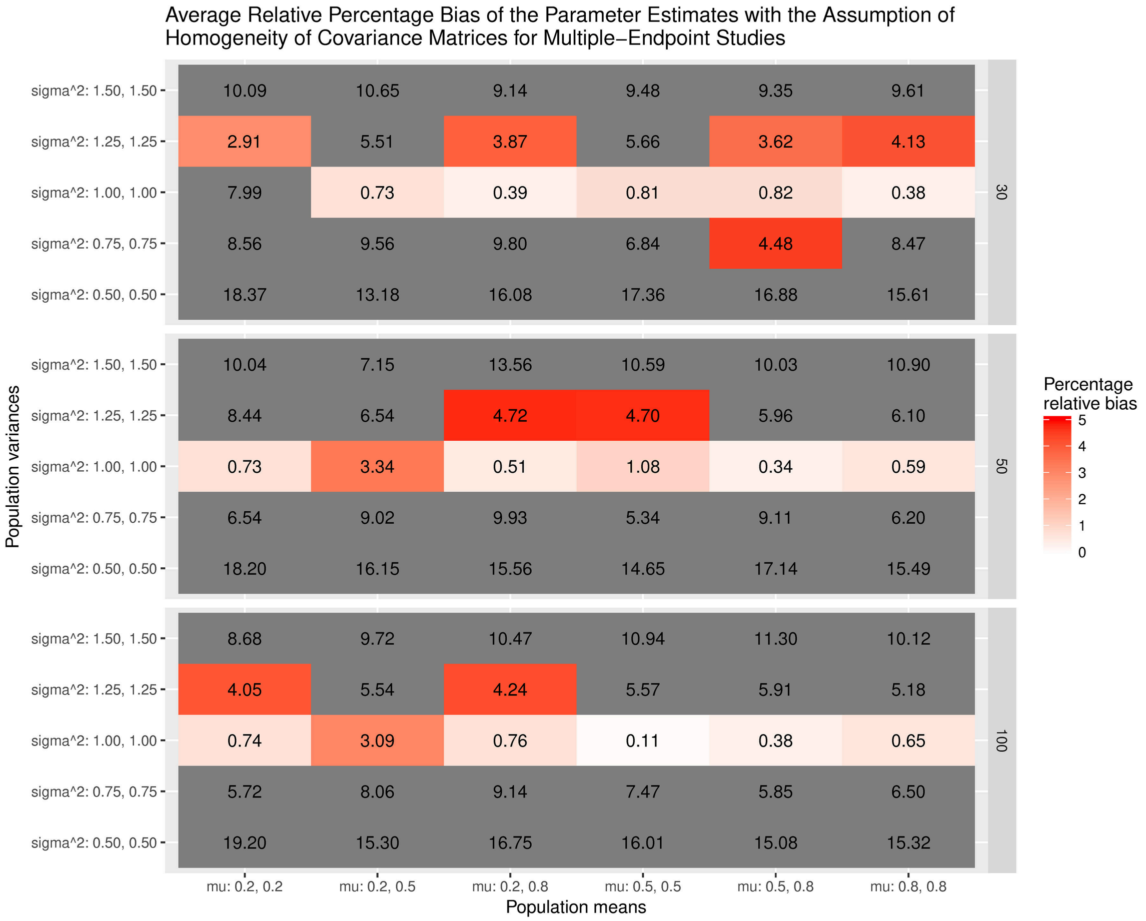


FIGURE 7 | Relative bias of the average of the parameter estimates for the multiple-endpoint studies with the assumption of homogeneity of covariance matrices.

covariance. Hoogland and Boomsma (1998) suggested that a proper estimation method should have a relative percentage bias of 10% on the *SE*. That is, the estimated *SE* should be within 0.90–1.1 of the empirical *SD* of $\hat{\theta}$. As we were using the sampling variance (SE^2, not *SE*), we used $(1.1^2 - 1) \approx 20\%$ as an indicator of good performance in estimating the sampling covariance matrix.

In the review process, one reviewer suggested displaying the individual parameter estimates $\hat{\theta}_1$ and $\hat{\theta}_2$. Due to space constraints, we put these results of the multiple-treatment studies in **Supplementary Materials 1**. Moreover, the same reviewer also suggested checking the performance under unbalanced sample sizes. We reran the simulation studies by introducing unbalanced sample sizes. The levels of sample sizes for the control, treatment 1, and treatment 2 groups were (100, 30, 50), (100, 50, 30), (30, 100, 50), (30, 50, 100), (50, 100, 30), and (50, 30, 100). The other factors were identical to the previous simulations. The results of the multiple-treatment and multiple-endpoint studies are shown in **Supplementary Materials 2**.

Results

The results were summarized in the heat maps, which provide an easy way to visualize the performance of the statistics. The x- and y-axes represent the population means and population variances separated by the sample sizes. A lighter color indicates a smaller bias than values with a darker color. When the bias is larger than the cut-off point (5% for the mean and 20% for the sampling variances or covariances), the color becomes gray.

Figures 3,4 show the relative bias of the effect sizes with and without the assumption of homogeneity of variances in calculating the effect sizes, respectively. One interesting finding was that the estimated effect size was generally unbiased regardless of whether or not the homogeneity of variances was assumed in the calculations. One speculation is that the average variances of the control group, which are always 1, and those of the treatment groups, at (0.75, 1.25) and (0.5, 1.5), are very close to 1. When these common *SD*s are used as the standardizers, the calculated effect sizes are still unbiased. The bias shrinks when the sample size gets bigger.

Figure 5 displays the relative bias of the sampling variances and covariances when the variances are assumed to be homogeneous when the effect sizes are estimated. The findings show that the sampling variances and covariances are unbiased only when the variances are actually homogeneous. When the population variances are heterogeneous, the sampling variances

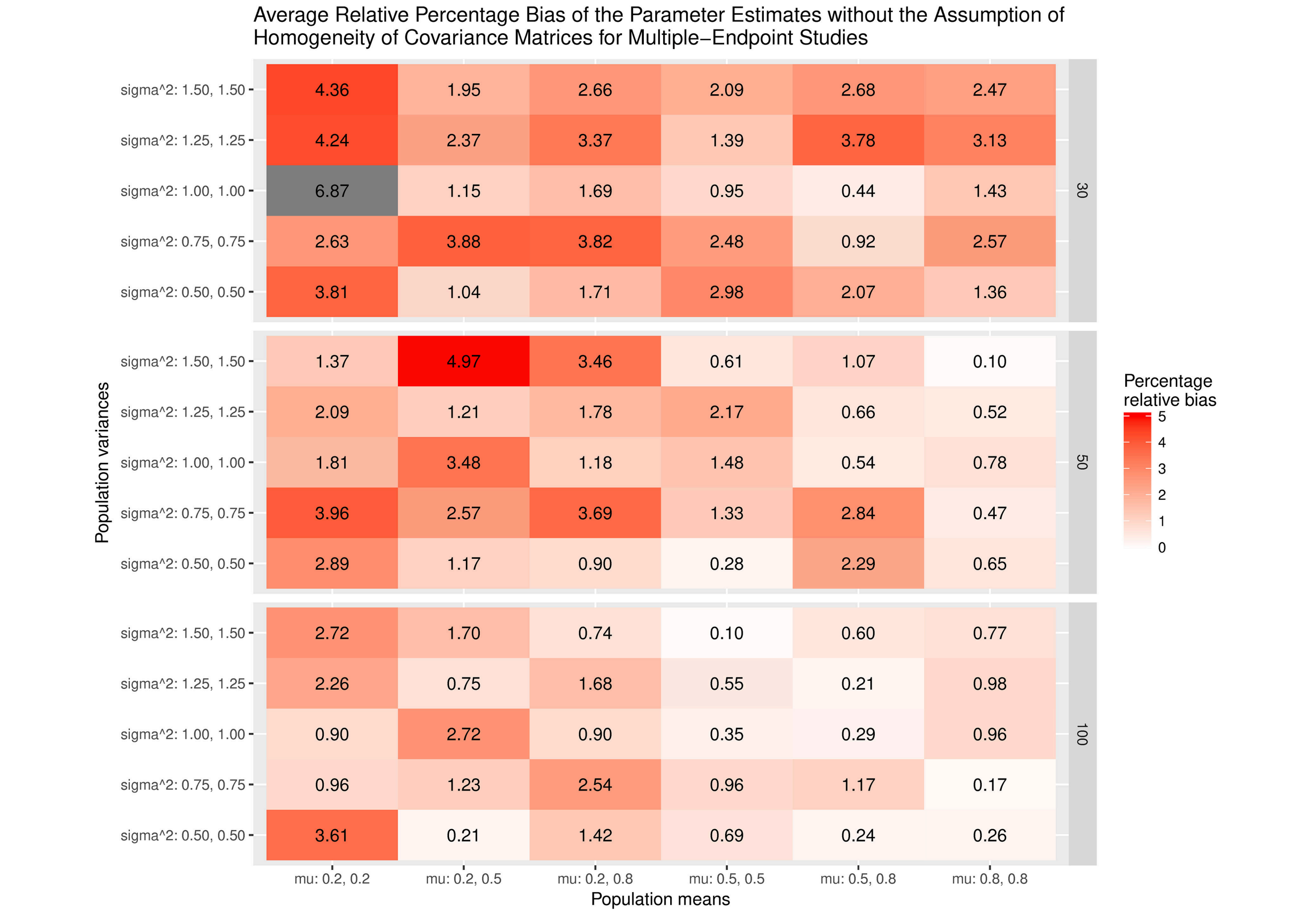


FIGURE 8 | Relative bias of the average of parameter estimates for the multiple-endpoint studies without the assumption of homogeneity of covariance matrices.

and covariances are biased. The most substantial bias occurs when the population variances have the largest differences (sigma∧2: 0.5, 1.5). **Figure 6** shows the relative bias of the sampling variances and covariances when the variances are not assumed to be homogeneous when estimating the effect sizes. In general, the bias is minimal, with the largest being only 12.6.

As a whole, the findings indicate that the effect sizes for the multiple treatment studies are estimated to be unbiased regardless of whether or not the homogeneity of variances is assumed in the calculations, given that the average of the treatment group variances are similar to that of the control group variance. However, the sampling variances and covariances are likely biased when the population variances are heterogeneous.

The patterns for the individual parameters in **Supplementary Material 1** are similar to those of the average parameters. Therefore, we will not reproduce them here. Regarding the simulation results of the unbalanced sample sizes in **Supplementary Material 2**, the estimated effect sizes with the homogeneity assumption are unbiased when the sample sizes in the control group are large (100, 30, 50) and (100, 50, 30). However, the bias of the estimated effect sizes with the homogeneity assumption becomes larger when the sample size of the control group is small, and the sample sizes in the treatment groups are unbalanced. The bias of the estimated effect sizes without the homogeneity assumption is generally small. Regarding the sampling variances and covariance, they are generally biased with the assumption of homogeneity, whereas they are generally unbiased without the assumption of homogeneity.

Study 2: Multiple-Endpoint Studies

The design was similar to those in the multiple-treatment studies. Two effect sizes were used in the simulation study, with one control group and one intervention group.

Methods

The population means and variances of the control group were fixed at 0 and 1, respectively, for reference. The population correlation between these two outcomes was set at 0.3, which is considered moderate in psychological research.

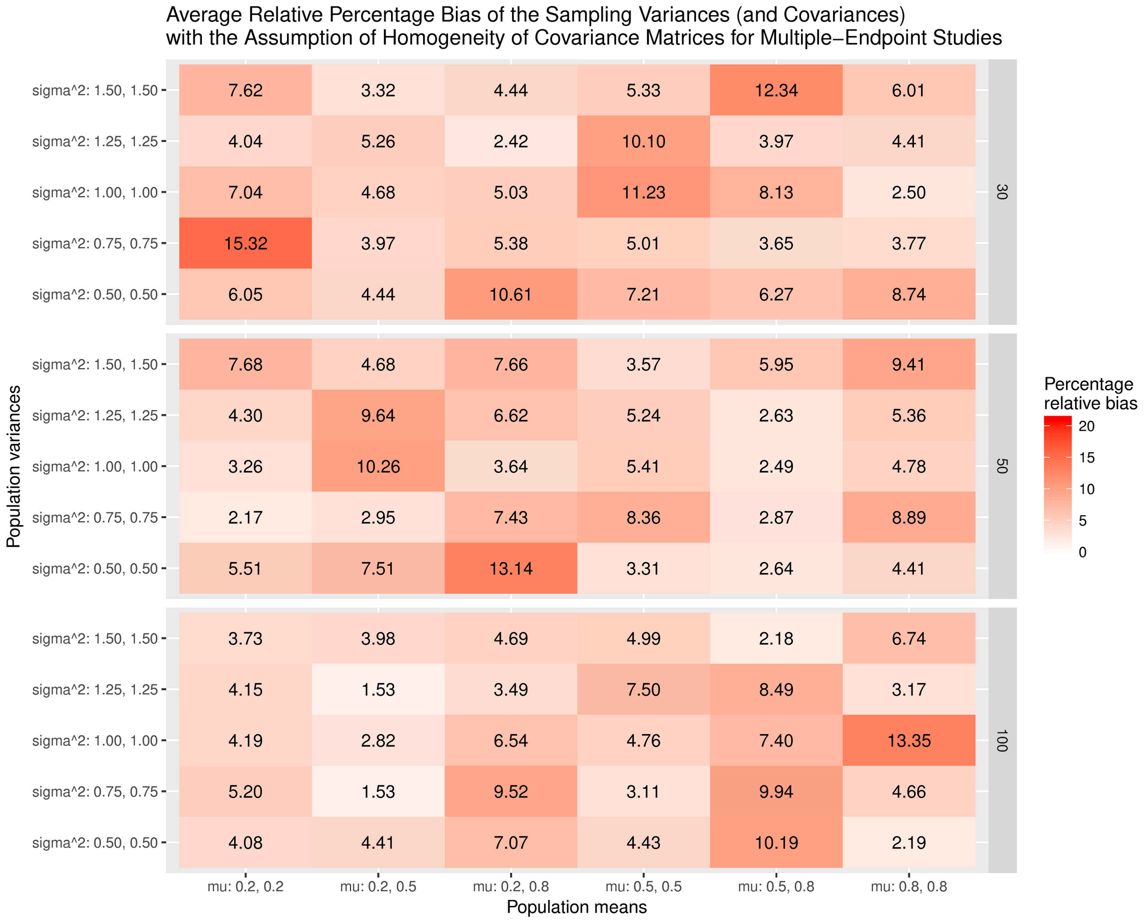


FIGURE 9 | Relative bias of the average of the sampling variances and covariance for the multiple-endpoint studies with the assumption of homogeneity of covariance matrices.

Population means

Six levels were used in the simulation study. The means for the two outcome variables in the intervention group were (0.2, 0.2), (0.2, 0.5), (0.2, 0.8), (0.5, 0.5), (0.5, 0.8), and (0.8, 0.8).

Population variances

Five levels for the intervention group were selected for the simulation. They were (1, 1), (0.5, 0.5), (0.75, 0.75), (1.25, 1.25), and (1.5, 1.5). When the population variance of the intervention group is (1, 1), the homogeneity of covariance matrices between studies is assumed; the assumption of the homogeneity of variances does not hold in the population.

Sample sizes

The design was assumed to be balanced. Three levels of sample sizes were selected, namely, 30, 50, and 100.

Therefore, there were a total of 6 × 5 × 3 = 90 conditions. One thousand replications were repeated for each condition.

Assessment of the empirical performance

The assessment was the same as those used in multiple-treatment studies. The average of the relative percentage bias of the effect size was used to evaluate the bias of the effect size. The average of the relative percentage bias of the sampling variances and covariances was used to assess the bias of the sampling covariance matrices. In the heat maps, 5 and 20% were used as the cutoff points.

Similar to the simulation studies in the multiple-treatment studies, we followed the advice of one reviewer by displaying the results of the individual effect sizes. The results are shown in **Supplementary Materials 3**. We also reran the simulation by introducing unbalanced sample sizes. The levels of the sample sizes in the control and treatment groups were (100, 30), (100, 50), (30, 100), (30, 50), (50, 100), and (50, 100). The results are displayed in **Supplementary Materials 4**.

Results

Figure 7 displays the average bias of the effect sizes when we assume the homogeneity of covariance matrices in calculating the effect sizes. When the covariance matrices are homogeneous (sigma∧2 = 1.00, 1.00), the effect sizes are generally unbiased except when mu = 0.2, 0.2 and the sample size = 30. However, the effect sizes are always biased when the covariance matrices

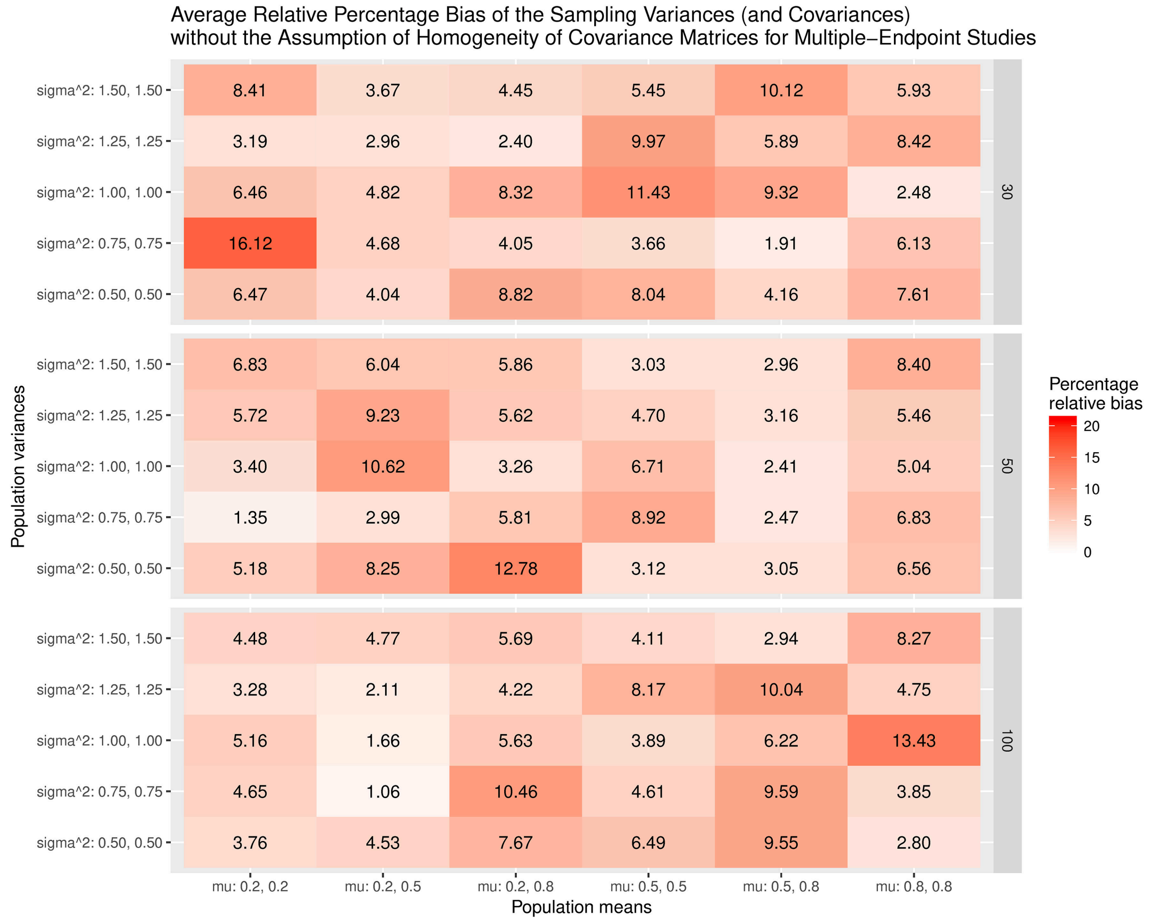


FIGURE 10 | Relative bias of the average of the sampling variances and covariance for the multiple-endpoint studies without the assumption of homogeneity of covariance matrices.

are not homogeneous. The most substantial relative bias can be up to 19%. This is expected because the variances of the treatment groups are very different from those of the control groups. **Figure 8** shows the average bias of the effect sizes when we do not assume the homogeneity of covariance matrices when calculating the effect sizes. The effect sizes are generally unbiased except when mu = 0.2, 0.2 and the sample size = 30.

Figure 9 displays the relative bias of the sampling variances and covariances when the effect sizes are estimated with the assumption of the homogeneity of covariance matrices. The bias is all below 20%. However, it should be noted that the effect sizes are biased. Thus, the results are still misleading. **Figure 10** shows the relative bias of the sampling variances and covariances when the effect sizes are estimated without the assumption of the homogeneity of covariance matrices. As can be seen, they are generally unbiased.

The patterns of the individual parameters displayed in **Supplementary Materials 3** are similar to those of the average parameters; therefore, we do not reproduce them here. Regarding the unbalanced data, the patterns are similar to those in multiple-treatment studies. The bias of the estimated effect sizes with the homogeneity assumption is much larger than that for the balanced data. On the other hand, the impact of the unbalanced sample sizes on the estimated effect sizes without the assumption of homogeneity is minimal.

To summarize, the estimated effect sizes are quite sensitive to the assumption of the homogeneity of covariance matrices. If the data are not homogeneous in covariance matrices and we incorrectly assume that they are, the estimated effect sizes are likely to be biased. On the other hand, the sampling covariance matrices are generally similar regardless of whether or not we have imposed the assumption of the homogeneity of covariance matrices.

CONCLUSION

This study shows that multivariate effect sizes for multiple-treatment and multiple-endpoint studies can easily be obtained using the SEM approach. Researchers may impose equality constraints on the variances and covariances, and the SEM packages will report the effect sizes and their sampling covariance matrices.

For multiple-treatment studies, the estimated effect sizes are unbiased regardless of whether or not we assume that the variances are homogeneous when calculating the effect sizes when the common *SD*s are close to the *SD*s of the control group. We may expect that there will be substantial bias when the common *SD*s are different from the *SD*s of the control group. Moreover, the estimated sampling covariance matrices are biased when the variances are heterogeneous, but we incorrectly assume that the variances are homogeneous.

For multiple-endpoint studies, the estimated effect sizes are biased when the covariance matrices are different, but we mistakenly assume that the covariance matrices are homogeneous. On the other hand, the sample covariance matrices are similar regardless of whether or not we have imposed the assumption of the homogeneity of covariance matrices when estimating the effect sizes.

The findings indicate that researchers should always check the assumptions before calculating the effect sizes. Researchers may also check the robustness of the findings by dropping these assumptions. By comparing the results with and without the assumption of the homogeneity of variances or covariance matrices, researchers may have a better idea of whether their substantive findings depend on these assumptions. Based on the simulation studies, it can be seen that the results are similar for the approaches with and without the assumption of the homogeneity of variances (or covariance matrices) when the data actually have the same variances (or covariance matrices). Therefore, the loss of efficiency from dropping the assumption of the homogeneity of variances (or covariance matrices) is small.

It should be noted that only a few factors were studied in the simulation studies. Further simulation studies may address the question of whether the findings are consistent in other conditions such as in those of unbalanced data and data with non-normal distributions. Another possible direction of research is to study how the assumption of the homogeneity of variances or covariance matrices impacts the actual parameter estimates in a meta-analysis. Such a study may provide stronger evidence to guide researchers on the issue of whether or not to report effect sizes with the assumption of homogeneity.

To conclude, it seems reasonable not to assume the homogeneity of variances (or covariance matrices) when calculating effect sizes for multiple-treatment and multiple-endpoint studies. The SEM approach provides a convenient device to calculate these effect sizes.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and approved it for publication.

FUNDING

This research was supported by the Academic Research Fund Tier 1 (FY2013-FRC5-002) from the Ministry of Education, Singapore.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2018.01387/full#supplementary-material

REFERENCES

Bond, C. F. Jr., Wiitala, W. L., and Dan, F. (2003). Meta-analysis of raw mean differences. *Psychol. Methods* 8, 406–418. doi: 10.1037/1082-989X.8.4.406

Borenstein, M., Hedges, L. V., Higgins, J. P. T., and Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester; Hoboken, NJ: John Wiley and Sons.

Card, N. A. (2012). *Applied Meta-Analysis for Social Science Research*. New York, NY: The Guilford Press.

Cheung, M. W. L. (2013). Multivariate meta-analysis as structural equation models. *Struc. Equ. Modeling* 20, 429–454. doi: 10.1080/10705511.2013.797827

Cheung, M. W. L. (2015a). *Meta-Analysis: A Structural Equation Modeling Approach*. Chichester: John Wiley and Sons, Inc.

Cheung, M. W. L. (2015b). metaSEM: an R package for meta-analysis using structural equation modeling. *Front. Psychol.* 5:1521. doi: 10.3389/fpsyg.2014.01521

Cheung, M. W. L., and Chan, W. (2004). Testing dependent correlation coefficients via structural equation modeling. *Org. Res. Methods* 7, 206–223. doi: 10.1177/1094428104264024

Cheung, M. W.-L., and Chan, W. (2005). Meta-analytic structural equation modeling: a two-stage approach. *Psychol. Methods* 10, 40–64. doi: 10.1037/1082-989X.10.1.40

Cohen, J. (1994). The earth is round ($p < .05$). *Am. Psychol*. 49, 997–1003. doi: 10.1037/0003-066X.49.12.997

Cumming, G. (2014). The new statistics: why and how. *Psychol. Sci.* 25, 7–29. doi: 10.1177/0956797613504966

Glass, G. V., McGaw, B., and Smith, M. L. (1981). *Meta-Analysis in Social Research*. Beverly Hills, CA: Sage Publications.

Gleser, L. J., and Olkin, I. (1994). "Stochastically dependent effect sizes," in *The Handbook of Research Synthesis,* eds H. Cooper and L. V. Hedges (New York, NY: Russell Sage Foundation), 339–355

Gleser, L. J., and Olkin, I. (2009). "Stochastically dependent effect sizes," in *The Handbook of Research Synthesis and Meta-analysis, 2nd Edn*, eds H. Cooper, L. V. Hedges, and J. C. Valentine (New York, NY: Russell Sage Foundation), 357–376).

Hoogland, J. J., and Boomsma, A. (1998). Robustness studies in covariance structure modeling an overview and a meta-analysis. *Soc. Methods Res. 26*, 329–367. doi: 10.1177/0049124198026003003

Hunter, J. E., and Hamilton, M. A. (2002). The advantages of using standardized scores in causal analysis. *Human Commun. Res.* 28, 552–561. doi: 10.1111/j.1468-2958.2002.tb00823.x

Jackson, D., Riley, R., and White, I. R. (2011). Multivariate meta-analysis: potential and promise. *Stat. Med*. 30, 2481–2498. doi: 10.1002/sim.4172

Kelley, K., and Preacher, K. J. (2012). On effect size. *Psychol. Methods* 17, 137–152. doi: 10.1037/a0028086

Kim, R.-S., and Becker, B. J. (2010). The degree of dependence between multiple-treatment effect sizes. *Multivar. Behav. Res*. 45, 213–238. doi: 10.1080/00273171003680104

Nam, I., Mengersen, K., and Garthwaite, P. (2003). Multivariate meta-analysis. *Stat. Med*. 22, 2309–2333. doi: 10.1002/sim.1410

Preacher, K. J. (2006). Testing complex correlational hypotheses with structural equation models. *Struct. Equ. Modeling.* 13, 520–543. doi: 10.1207/s15328007sem1304_2

R Development Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna. Available online at: http://www.R-project.org/

Raykov, T. (2001). Testing multivariable covariance structure and means hypotheses via structural equation modeling. *Struct. Equ. Modeling*. 8, 224–256. doi: 10.1207/S15328007SEM0802_4

Schmidt, F. L., and Hunter, J. E. (2015). *Methods of Meta-Analysis: Correcting Error and Bias In Research Findings, 3rd Edn*. Thousand Oaks, CA: Sage.

Thompson, C. G., and Becker, B. J. (2014). The impact of multiple endpoint dependency on Q and I2 in meta-analysis. *Res. Synthesis Methods* 5, 235–253. doi: 10.1002/jrsm.1110

Wilkinson, L., and Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *Am. Psychol.* 54, 594–604. doi: 10.1037/0003-066X.54.8.594

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer, JP, declared a past co-authorship with the author, MC, to the handling Editor.

METHODS
published: 28 July 2017
doi: 10.3389/fpsyg.2017.01286

Evaluating Fit Indices for Multivariate *t*-Based Structural Equation Modeling with Data Contamination

Mark H. C. Lai and Jiaqi Zhang*

School of Education, University of Cincinnati, Cincinnati, OH, United States

In conventional structural equation modeling (SEM), with the presence of even a tiny amount of data contamination due to outliers or influential observations, normal-theory maximum likelihood (ML-Normal) is not efficient and can be severely biased. The multivariate-*t*-based SEM, which recently got implemented in Mplus as an approach for mixture modeling, represents a robust estimation alternative to downweigh the impact of outliers and influential observations. To our knowledge, the use of maximum likelihood estimation with a multivariate-*t* model (ML-*t*) to handle outliers has not been shown in SEM literature. In this paper we demonstrate the use of ML-*t* using the classic Holzinger and Swineford (1939) data set with a few observations modified as outliers or influential observations. A simulation study is then conducted to examine the performance of fit indices and information criteria under ML-Normal and ML-*t* in the presence of outliers. Results showed that whereas all fit indices got worse for ML-Normal with increasing amount of outliers and influential observations, their values were relatively stable with ML-*t*, and the use of information criteria was effective in selecting ML-normal without data contamination and selecting ML-*t* with data contamination, especially when the sample size was at least 200.

Keywords: structural equation modeling, robustness, outliers, data contamination, fit indices

OPEN ACCESS

Edited by:
Mike W.-L. Cheung,
National University of Singapore,
Singapore

Reviewed by:
Ji Yeh Choi,
McGill University, Canada
Victoria Savalei,
University of British Columbia, Canada

****Correspondence:***
Mark H. C. Lai
mark.lai@uc.edu

Specialty section:
This article was submitted to Quantitative Psychology and Measurement,
a section of the journal Frontiers in Psychology

Received: *23 April 2017*
Accepted: *13 July 2017*
Published: *28 July 2017*

Citation:
Lai MHC and Zhang J (2017) Evaluating Fit Indices for Multivariate t-Based Structural Equation Modeling with Data Contamination. Front. Psychol. 8:1286. doi: 10.3389/fpsyg.2017.01286

Although identification of outliers and influential observations is a standard practice in regression models, less attention has been given to such issues in structural equation modeling (SEM) (Pek and MacCallum, 2011). Nevertheless, as pointed out in previous literature (e.g., Yuan and Bentler, 2001; Yuan and Zhong, 2013), normal-theory based SEM is not robust to data contamination, and a small proportion of outliers and influential observations can bias parameter estimation, the likelihood ratio test statistic (LRT; also commonly referred to as the model χ^2), and fit indices based on LRT. While robust modeling by replacing the normality assumption with one that assumes the error terms follow a heavier-tailed *t* distribution has long been discussed in regression models and multilevel models (e.g., Pinheiro et al., 2001; Gelman and Hill, 2006), not until recently are multivariate-*t* SEM models accessible to researchers, and there has been very little research on the usefulness of such models in the presence of data contamination. In this paper we first provide brief background information on the use of the multivariate *t* distribution for robust SEM modeling. We then demonstrate with a real data set how five outlying cases can have a severe impact on model fit and parameter estimates under normal-theory SEM (ML-Normal), and show that estimation using a *t* model (ML-*t*) produces similar inferences with and without data contamination. Finally, we conduct a simulation study to evaluate the performance of commonly used fit indices of ML-Normal, ML-*t*, and the use of Huber-type weights with and without data contamination, and the effectiveness of information criteria in selecting between ML-Normal and ML-*t*, across conditions of model misspecifications, sample sizes, and proportions of outliers and influential observations.

OUTLIERS AND INFLUENTIAL OBSERVATIONS

Whereas topics related to outliers, or more generally *data contamination*, are commonly discussed in quantitative research methodology textbooks, in practice researchers do not always agree on their definitions and how best to handle them. For instance, in a review of organizational research, Aguinis et al. (2013) found 14 different definitions of outliers (which include but are not limited to cases with high leverage and with large influence on parameter estimates and model fit) and 20 different ways to handle them. Also, outliers of different nature require different treatments, and Aguinis et al. summarized the definitions of outliers in three categories: (a) those due to correctable errors such as input error, (b) those exhibiting idiosyncratic characteristics and of interest themselves (c) those exerting disproportionately large influence on the substantive conclusion regarding a model of interest. In this paper we focus on robust inference in SEM with data contamination in category (c).

Following Pek and MacCallum (2011), in this study we distinguish between *outliers* and *influential observations* for SEM, as they may have differential impacts on parameter estimation and model fit indices. Outliers are cases that lie far away from most other data points. In regression with only one response variable, outliers are cases with a large deviation from its predicted value based on the regression line. In multivariate analyses such as SEM, the distance of an observation from the center of most of the data points is commonly quantified by the Mahalanobis distance (d), where:

$$d_i = \sqrt{(\mathbf{y}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}(\mathbf{y}_i - \boldsymbol{\mu})}, \tag{1}$$

$\mathbf{y}_i = [y_{1i}, \ldots, y_{ki}]$ is the data vector for the ith observation on p observed variables, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the mean vector and the covariance matrix of the p observed variables. On the other hand, influential observations are those that exert large influence on model fit and parameter estimation; in other words, parameter estimates and model fit indices will show relatively large changes when the influential cases are removed. Despite the conceptual differences between outliers and influential observations, they are not mutually exclusive as some outliers can also exert strong influence on the results.

Although in this paper we focus on methods to obtain fit indices and parameter estimates that are robust to extreme observations without necessarily identifying the outliers and influential observations, researchers are generally recommended to carefully inspect observations and identify correctable data entry errors and truly idiosyncratic observations. Examples of techniques for identifying outliers and influential observations in SEM were Cook's distance, Mahalanobis distance, and likelihood distance. Readers can consult Aguinis et al. (2013), Pek and MacCallum (2011), and Yuan and Zhang (2012) for more in-depth discussions on tools and procedures for identifying outliers and influential observations.

Here we borrow the notations from Asparouhov and Muthén (2015) for the linear SEM model and define outliers and influential observations as discussed in Yuan and Zhong (2008, 2013). For a model with p observed variables measuring q latent variables, we assume that the observed p-variate observed variable $\mathbf{Y}$ has a measurement model:

$$\mathbf{Y} = \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\eta} + \mathbf{e}, \tag{2}$$

where $\boldsymbol{\nu}$ is a $p \times 1$ vector of measurement intercepts, $\boldsymbol{\Lambda}$ is a $p \times q$ factor loading matrix, and $\mathbf{e}$ is a $p \times 1$ random vector containing measurement error terms. $\boldsymbol{\eta}$ is a q-variate latent variable with a structural model:

$$\boldsymbol{\eta} = \boldsymbol{\alpha} + \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\mathbf{X} + \boldsymbol{\xi}, \tag{3}$$

where the effects among latent factors were captured by $\mathbf{B}$, the effects of exogenous covariates $\mathbf{X}$ were captured by $\boldsymbol{\Gamma}$, $\boldsymbol{\xi}$ is a random vector of disturbance terms, and $\boldsymbol{\alpha}$ contains the latent regression intercepts. It is common to impose the normality assumption such that the joint distribution of $\mathbf{e}$ and $\boldsymbol{\xi}$ is multivariate normal, with:

$$(\mathbf{e}, \boldsymbol{\xi}) \sim N_{p+q}\left(\mathbf{0}, \begin{bmatrix} \boldsymbol{\Theta} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Psi} \end{bmatrix}\right). \tag{4}$$

As discussed in Yuan and Zhong (2008), outliers in SEM have large values of $\mathbf{e}$, and will inflate the covariance matrix of the outcome variables $\boldsymbol{\Sigma}$. However, it may or may not have large values in $\boldsymbol{\eta}$. On the other hand, influential observations have extreme values in $\boldsymbol{\xi}$ and will inflate $\boldsymbol{\Psi}$ and also $\boldsymbol{\Sigma}$. Influential observations can be good or bad: good influential observations have extreme $\boldsymbol{\xi}$ but not extreme $\mathbf{e}$ values, and will not negatively impact model fit as it is not considered outliers; bad influential observations, on the other hand, have both extreme $\boldsymbol{\xi}$ and $\mathbf{e}$ values, and will negatively impact model fit.

Impact of Outliers and Influential Observations

Under the normal model, a very small portion of outliers and influential observations can have a huge impact on parameter estimates and model fit. For example, Yuan and Bentler (2001) showed mathematically that existence of outliers can greatly inflate the Type I error rates of LRT and related test statistics adjusting for non-normality under ML-Normal, and the LRT statistic could be inflated by more than five times in values in an example given in Yuan and Zhong (2008); Yuan and Zhong (2008) also showed that in confirmatory factor analysis (CFA), about 3% of outliers could substantially bias the factor loading estimates by more than 50% and inflate the latent factor variance and covariance estimates by 3–10 times, whereas 3% of bad influential observations could produce even greater biases on all parameter estimates in CFA. Yuan and Zhong (2013) showed mathematically and illustrated with modified real data sets that outliers lead to worse fit indices such as RMSEA and CFI, whereas bad influential observations can lead to worse RMSEA but also *better* CFI in some situations. In summary, a few outliers and bad influential observations can lead to biased and inefficient parameter estimates and produced misleading and sometime contradictory information about model fit.

Despite the documented impact of outliers and influential observations, detection and diagnostics of such observations were rarely performed and reported in real research, and the use of SEM methods that are robust to data contamination has been scarce. For example, Aguinis et al. (2013) reviewed 232 methodological and substantive journal articles in organizational science journals that addressed issues about outliers and influential observations, and only five of them were related to SEM, despite the popularity of SEM in the past two decades. One possible reason is that practical guidelines on handling outliers and influential observations were developed more recently (e.g., Pek and MacCallum, 2011; Aguinis et al., 2013). Another possible reason is that existing methods for detecting and handling outliers and influential observations in SEM require researchers to use specialized programs (e.g., Sterba and Pek, 2012; Yuan and Zhang, 2012), thus creating additional burden for researchers if they are not familiar with those programs.

Existing Robust Estimation Methods in SEM

Because a small proportion of outliers and bad influential observations can produce invalid assessment of model fit and parameter estimates, it is important to have methods that produce consistent and efficient estimation and give robust model fit information in the presence of data contamination, assuming that those extreme values are not due to correctable errors (e.g., data entry errors) and the goal is to obtain inferences based on the majority of the sample. One such method in SEM is to replace the squared loss function in estimating the mean vector and covariance matrix by one that downweighs cases exerting unproportionally large influence on the model (Yuan and Bentler, 1998a,b, 2000; Yuan et al., 2000, 2004), which has been commonly used in robust regression. One popular choice of the weight function is the Huber-type weight function, which replaces the squared loss function by a linear function when the Mahalanobis distance of a case exceeds a prespecified cutoff, u, where u^2 is the $(1 - \varphi)$th quantile of a chi-square distribution. In other words, φ is the theoretical proportion of cases to be downweighed under normality with no data contamination.

In the two-stage robust method (TSR; Yuan and Bentler, 1998a), one first obtains robust means and covariances estimates with Huber-type weights:

$$\hat{\boldsymbol{\mu}}_{\text{TSR}} = \frac{\sum_{i=1}^{n} w_1(d_i)\mathbf{y}_i}{\sum_{i=1}^{n} w_1(d_i)},$$

$$\hat{\boldsymbol{\Sigma}}_{\text{TSR}} = \frac{1}{n}\sum_{i=1}^{n} w_2(d_i)(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_{\text{TSR}})(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_{\text{TSR}})^{\top},$$

where d_i is the Mahalanobis distance of the ith observation as defined in (1), and the Huber-type weights are:

$$w_1(d) = \begin{cases} 1 & \text{if } d \le u \\ u/d & \text{if } d > u \end{cases},$$

$$w_2(d) = [w_1(d)]^2/\tau,$$

and τ is a constant to ensure that $\hat{\boldsymbol{\Sigma}}_{\text{TSR}}$ is unbiased for $\boldsymbol{\Sigma}$. $\hat{\boldsymbol{\mu}}_{\text{TSR}}$ and $\hat{\boldsymbol{\Sigma}}_{\text{TSR}}$ can then be input into common SEM software to obtain robust parameter estimates. However, with TSR, the LRT, fit indices, and standard errors reported in standard software outputs cannot be used; instead, one needs to compute those according to the formulas given in Yuan and Bentler (1998b), which are also available in R using the `rsem` package (Yuan and Zhang, 2015). An alternative is the direct robust method (Yuan and Zhong, 2008), which uses iteratively reweighted least squares to obtain the Mahanobis distance of each observation based on $\hat{\mathbf{e}}_i$, downweighs the cases with large Mahanalobis distance, and re-estimates the other model parameters until convergence. For more discussions on TSR and the direct robust method, please consult Yuan and Zhong (2008) and Yuan and Hayashi (2010).

Previous work has shown, mathematically and through empirical examples and simulation studies, that TSR and the direct robust method provided less biased parameter estimates with smaller sampling variability (i.e., greater efficiency) and adjusted LRT relatively insensitive to data contamination. Yuan and Zhong (2013) also demonstrated that TSR and the direct robust method provided fit indices closer to the population value with no outliers or influential observations. However, there are two drawbacks of using Huber-type weights, including (a) the need to select a tuning parameter that determines the proportion of observations being downweighed, and (b) the difficulty in obtaining likelihood-based information criteria for model selection. Therefore, the multivariate-*t* model implemented in Mplus (Muthén and Muthén, 1998–2015), which implicitly also has the effect of downweighing extreme cases but solves the difficulties (a) and (b), will be an attractive alternative to some researchers given its ease of use.

Multivariate-*t* Based SEM

An alternative to handle outliers in regression is to replace the normality assumption on the error terms by a heavy-tailed distribution, where the heavier tails reduce the impact of extreme cases on inferences of the center and variability of the data. One common choice of heavy-tailed distributions is the Student's *t* distribution (Zellner, 1976), with a degree of freedom (*df*) parameter controlling the tail density; a smaller *df* put more weight on the tails, whereas a $df > 30$ effectively makes the *t* distribution closely match the normal distribution. Such a class of models has long been discussed in regression (Gelman and Hill, 2006) and in multilevel modeling (Pinheiro et al., 2001).

In SEM, estimation involving the multivariate *t* distribution is not new. Indeed, as stated in Yuan and Bentler (1998b), by using a specific weighting scheme of the observations in a way analogous to the Huber-type weights, one can obtain robust estimates of mean vector and covariance matrix for the observed variables as the maximum likelihood estimates based on a multivariate t-distribution. Yuan and Bentler (1998b) and Yuan et al. (2004) showed that using robust covariances based on the multivariate t distribution also performed well in terms of providing less biased parameter estimates and more robust LRT results. However, the procedure discussed in Yuan and Bentler (1998b) required pre-defined degrees of freedom parameter and had not been implemented in statistical software.

Recently, a multivariate-t model for SEM has been incorporated into the software Mplus, together with the skew-normal and skew-t family (Asparouhov and Muthén, 2015). The documented usage of such models is for mixture modeling with skewed and heavy-tailed compositions to avoid spurious latent classes, and to our knowledge there has been no discussion on using the t-based model for robust SEM in the presence of data contamination. The multivariate t distribution, $t_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, df)$, is a p-variate generalization of the Student's t distribution with a single df parameter, a location vector μ, and a scale matrix $\boldsymbol{\Sigma}$. The mean of the distribution is $\boldsymbol{\mu}$ and the covariance matrix is $[df/(df-2)]\boldsymbol{\Sigma}$ for $df > 2$. Therefore, when df is small, estimates of parameters in $\boldsymbol{\Sigma}$ in the t-based model has a different interpretation than those in the normal-based model.

Continuing from the model defined in Equations (2)–(4), with t-based SEM one simply replaces the distributional assumption in (4) by:

$$(\mathbf{e}, \boldsymbol{\xi}) \sim t_{p+q}\left(\mathbf{0}, \begin{bmatrix} \boldsymbol{\Theta} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Psi} \end{bmatrix}, df\right). \quad (5)$$

This is equivalent to the model with the conditional distribution $\mathbf{Y}|\mathbf{X} \sim t_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, df)$, where

$$\boldsymbol{\mu} = \boldsymbol{\nu} + \boldsymbol{\Lambda}(\mathbf{I} - \mathbf{B})^{-1}(\boldsymbol{\alpha} + \boldsymbol{\Gamma}\mathbf{X}), \quad (6)$$

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}(\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Psi}[\boldsymbol{\Lambda}(\mathbf{I} - \mathbf{B})^{-1}]^{\top}\boldsymbol{\Lambda}^{\top}. \quad (7)$$

As the model likelihood can be specified, maximum likelihood can be used to estimate all model parameters as well as df, thus avoiding the need to choose a tuning parameter as in using Huber-type weights. This also allows the computation of information criteria such as AIC (Akaike Information Criteria), BIC (Bayesian Information Criteria), and SABIC (sample-size adjusted BIC in Mplus).

Although previous studies have shown that the use of robust covariance matrix based on weights corresponding to a multivariate t distribution provided good parameter estimates and LRT statistics similar to those obtained without outliers under ML-Normal (Yuan and Bentler, 1998b), to our knowledge no analytic and simulation studies have evaluated the performance of LRT and fit indices obtained under the multivariate-t-based SEM as implemented in Mplus (i.e., ML-t), with df being estimated instead of specified by users. Given that parameter estimates are not trustworthy when the model fit is sup-optimal, accurate assessment of model fit for an SEM model is of paramount importance. Therefore, this study is an important first step in examining SEM models with ML-t as a robust option in handling outliers and influential observations.

It should be clarified that our discussion is limited to robust models that are insensitive to the influence of data contamination, which is different from SEM methods that are robust to non-normality, such as the corrections in LRT and standard errors proposed by Satorra and Bentler (1994). Although there are some commonalities between the two topics, the former focuses on downweighing extreme observations to obtain estimates and inferences when the normality assumption still holds approximately for the majority of the data, and robust SEM methods for non-normality focuses on obtaining inferences when the normality assumption is violated in general. Whereas the latter has received much attention in SEM literature (e.g., Bentler, 1983; Browne, 1984), they generally require estimation of some higher moments of the sample data, which would be highly unstable in the presence of data contamination. Indeed, as would be mentioned in the discussion, we found the Satorra-Bentler correction performed sub-optimally in the presence of data contamination.

REAL DATA DEMONSTRATION

We now briefly demonstrate the use of SEM with ML-t using the classic Holzinger and Swineford (1939) data set and a modified version where five observations were changed to have strong influence to model fit. The nine variables are cognitive test scores for 145 students. Using ML-Normal with the original data set, a 3-factor CFA model with a cross-loading of item 9 on factor 1 fit the data well, with $\chi^2(df{=}23, N{=}145){=}28.29$, $p{=}0.205$, RMSEA=0.040, CFI=0.989, SRMR=0.040. One can also use the `DIST=TDISTRIBUTION` option in Mplus to fit the same model with t-likelihood, and add the `OUTPUT: H1MODEL` option to obtain χ^2 statistic and fit indices. Using the same data, one obtains $\chi^2(df{=}23, N{=}145){=}25.29$, $p{=}0.335$, RMSEA=0.026, CFI=0.994, which were close to the values with ML-Normal. SRMR is not yet obtainable as it does not directly depend on χ^2. The estimated df using maximum likelihood is 24.5, with a 95% confidence interval (CI) of $[15.9, 34.1]$. Using TSR as implemented in the R package `rsem` with the same CFA model and data, we have $\chi^2(df{=}23, N{=}145){=}26.82$, $p{=}0.264$, RMSEA=0.034, CFI=0.992, SRMR=0.038. AIC, BIC, and SABIC all preferred ML-t over ML-Normal with differences of 8.5, 5.5, and 8.7 respectively, indicating some evidence for slightly heavier tails in the sample distributions even without modifications.[1] However, the fit information and parameter estimates (as shown in **Table 1**) under all three methods were similar, so the choice is trivial and one can be more confident that data contamination should not be an issue.

We then use a modified data set described in Yuan and Zhong (2013, p. 131, Data Set D4) containing five bad influential observations (i.e., 3%). The Mahanalobis distances

[1] It is still an open question how information criteria should be defined when using TSR, and whether the values obtained can be compared with information criteria obtained using maximum likelihood. Therefore, we did not report information criteria with TSR.

TABLE 1 | Parameter estimates and standard errors from the real data demonstration.

	ML-Normal		ML-*t*		TSR	
Parameters	**Original**	**Modified**	**Original**	**Modified**	**Original**	**Modified**
λ_{11}	1.00 (—)	1.00 (—)	1.00 (—)	1.00 (—)	1.00 (—)	1.00 (—)
λ_{21}	0.66 (0.14)	1.14 (0.06)	0.63 (0.14)	0.73 (0.13)	0.62 (0.13)	0.77 (0.14)
λ_{31}	0.84 (0.15)	0.43 (0.04)	0.82 (0.15)	0.74 (0.12)	0.78 (0.13)	0.73 (0.12)
λ_{42}	1.00 (—)	1.00 (—)	1.00 (—)	1.00 (—)	1.00 (—)	1.00 (—)
λ_{52}	0.99 (0.09)	1.47 (0.06)	1.02 (0.09)	1.09 (0.10)	1.04 (0.09)	1.07 (0.09)
λ_{62}	0.96 (0.08)	1.14 (0.05)	0.97 (0.09)	0.95 (0.08)	0.96 (0.09)	0.93 (0.08)
λ_{73}	1.00 (—)	1.00 (—)	1.00 (—)	1.00 (—)	1.00 (—)	1.00 (—)
λ_{83}	1.27 (0.23)	1.49 (0.06)	1.27 (0.24)	1.37 (0.20)	1.18 (0.19)	1.29 (0.16)
λ_{91}	0.56 (0.12)	1.18 (0.20)	0.55 (0.12)	0.56 (0.12)	0.51 (0.11)	0.52 (0.14)
λ_{93}	0.64 (0.14)	0.49 (0.21)	0.63 (0.14)	0.66 (0.13)	0.65 (0.13)	0.69 (0.14)
ψ_{11}	0.67 (0.16)	3.95 (0.55)	0.64 (0.16)	0.72 (0.16)	0.70 (0.16)	0.81 (0.21)
ψ_{22}	0.94 (0.15)	3.11 (0.42)	0.85 (0.15)	0.85 (0.15)	0.87 (0.14)	0.97 (0.16)
ψ_{33}	0.50 (0.13)	3.33 (0.47)	0.46 (0.13)	0.52 (0.13)	0.50 (0.13)	0.62 (0.14)

ML, maximum likelihood; TSR, Two-stage robust methods with Huber-type weights downweighing 10% of observations; λ, factor loading; ψ, factor variance. Standard errors were shown in parentheses.

of the five modified observations were between 4.22 and 128.26, compared to 1.41 to 24.84 for the other observations. ML-Normal gave $\chi^2(df{=}23, N{=}145){=}57.80$, $p < 0.001$, RMSEA=0.095, CFI=0.984, SRMR=0.020. Although both χ^2 and RMSEA indicated worse model fit, the impact on CFI was small and SRMR actually indicated better model fit. Therefore, with the presence of bad influential observations, ML-Normal gave ambiguous fit information. ML-*t*, on the other hand, gave $\chi^2(df{=}23, N{=}145){=}24.81$, $p{=}0.360$, RMSEA=0.023, CFI=0.996; TSR gave $\chi^2(df{=}23, N{=}145){=}30.55$, $p{=}0.134$, RMSEA=0.048, CFI=0.987, SRMR=0.037. Thus, the fit information using ML-*t* or TSR was comparable with or without the bad influential cases, and AIC, BIC, and SABIC all strongly favored ML-*t* over ML-Normal with differences in values of more than 300. The parameter estimates were shown in **Table 1**, which shows that whereas estimates were strongly affected by the influential observations using ML-Normal, they were robust with ML-*t* and TSR.

SIMULATION STUDY

We conducted a Monte Carlo simulation study to compare the performance of fit indices (χ^2, RMSEA, and CFI) under ML-Normal, ML-*t*, and TSR, as well as model comparisons between ML-Normal and ML-*t* using AIC, BIC, and SABIC. Note that SRMR is not available in Mplus 7.4 with ML-*t*. A 2 (type of data contamination) × 4 (proportion of outliers/influential observations) × 2 (model misspecification) × 3 (sample size) design was used, as described later. For all simulation conditions, data were generated first from a 3-factor 9-indicator model similar to the previous real data demonstration, which was also used in Zhong and Yuan (2011), with a cross-loading of item 9 on Factor 3. Specifically,

$$\mathbf{\Lambda} = \begin{bmatrix} 1.0 & 0.9 & 1.1 & 0 & 0 & 0 & 0 & 0 & 0.5 \\ 0 & 0 & 0 & 1.0 & 0.7 & 0.55 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1.3 & 1.25 \end{bmatrix}^{\top},$$

$$\mathbf{\Psi} = \begin{bmatrix} 1 & 0.25 & 0.25 \\ 0.25 & 1 & 0.25 \\ 0.25 & 0.25 & 1 \end{bmatrix},$$

and $\mathbf{\Theta} = \mathbf{I}$ is a 9 × 9 identity matrix. All factor means and intercepts were set to zero for simplification. For each simulation condition, we ran 2,000 replications.

Factor scores, $\boldsymbol{\xi} \sim N(\mathbf{0}, \mathbf{\Psi})$, and multivariate normal data, denoted as $\mathbf{y}^{(1)}$, were generated according to the above model using R 3.3.1 (R Core Team, 2016), and outliers or bad influential observations were introduced using methods described in Zhong and Yuan (2011). Let $\varepsilon = 0$, 0.05, 0.075, or 0.10 be the proportion of cases that are either outliers or bad influential observations. For conditions with outliers, εN observations were modified as $\mathbf{y}_i^{\text{mod}} = \mathbf{y}_i^{(1)} + h_i \mathbf{\Lambda}^o \boldsymbol{\xi}_i$, where $h_i \sim \exp(z_i)$ and z_is were randomly generated from independent $N(0, 1)$ distributions and varied across replications; $\mathbf{\Lambda}^o$ is the modified loading matrix such that $\mathbf{\Lambda}^\top \mathbf{\Lambda}^o = \mathbf{0}$ as discussed in Yuan and Zhong (2013). For conditions with bad influential observations, εN observations were modified as $\mathbf{y}_i^{\text{mod}} = h_i \mathbf{y}_i^{(1)}$. The sample size ($N$) was either 100, 200, or 500, as most studies using SEM had $N \geq 100$ (Jackson et al., 2009), and SEM was generally not recommended with a small sample size (Kline, 2011). The Mahalanobis distances for the outliers or influential observations varied across replications, and for each replication the maximum M-distance was 15.8–76.0, 19.7–149.5, 14.5–350.1 for $N = 100$, 200, and 500 with outliers (see **Figure 1** for an example), and 29.1–92.2, 30.3–190.4, 26.7–472.5 for $N = 100$, 200, and 500 with bad influential observations. For each generated data set, we

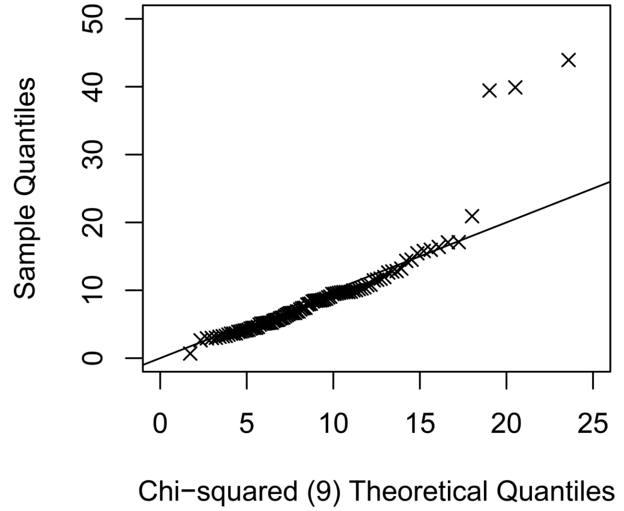


FIGURE 1 | An example quantile-quantile plot of the sample Mahanalobis distance against the theoretical values of a chi-squared distribution with 9 degrees of freedom for the condition with $N = 100$ and 5 outliers. Multivariate kurtosis = 9.66 for this simulated data set.

either fit a correctly specified model or a misspecified model with no cross-loading (with population RMSEA = 0.061, population CFI = 0.96), and for each model we used either ML-Normal or ML-*t* and obtain fit indices in Mplus, and used `rsem` to obtain fit indices with TSR (10% observations downweighed).

Evaluation Criteria

For each condition, we evaluated the rejection rates of LRT at $p < 0.05$ using the three methods, with rates close to 5% being optimal for conditions with a correctly specified model, and rates close to the empirical power with no data contamination best for conditions with a misspecified model. For RMSEA and CFI we graphically examine the distributions of the sample fit values, with preference given to methods providing sample fit values close to population fit values and with small variabilities across replications. Finally, we examined the empirical probability of information criteria selecting ML-*t* over ML-Normal (i.e., having smaller values for ML-*t*).

Simulation Results

Convergence

Convergence rates were above 90.5% for ML-Normal and above 93.3% for ML-*t* for conditions with correctly specified model, and were above 89.7% and above 87.8% for conditions with misspecified model. Convergence was better for conditions with $N \geq 200$ (> 98.3% for ML-*t* and > 92.5% for ML-Normal); it was worst for ML-Normal with 10% bad influential observations (90.5% for $N = 100$ and 92.8% for $N = 200$ for correctly specified model and 89.7% for $N = 100$ and 92.5% for misspecified model), and for ML-*t* with small sample size (93.3–95.2% for correctly specified model and 87.8–91.0% for misspecified model when $N = 100$). Convergence was generally better for ML-*t* than for ML-Normal in conditions with outliers or bad influential observations when $N \geq 200$. Convergence rates were at least 99.2% with TSR.

Fit Indices

As shown in **Figures 2–5**, the results of ML-*t* and TSR were almost identical, so in the following sections we mainly prsented results for ML-Normal and ML-*t*.

Outliers with correctly specified model

Figure 2 showed the boxplots of sample values of LRT, RMSEA, and CFI for conditions with a correctly specified model and the presence of outliers. With ML-Normal, LRT, RMSEA, and CFI became substantially worse and more variable with increasing proportion of outliers. When $N = 100$, median LRT increased slightly from 23.26 (adjusted median absolute deviation, $SD = 7.28$) with no outliers to 28.96 ($SD = 19.63$) with 10% outliers; when $N = 500$, median LRT increased dramatically from 22.73 ($SD = 6.82$) with no outliers to 51.27 ($SD = 71.74$) with 10% outliers. Empirical Type I error rates of LRT were inflated from 5.4 to 7.3% with no outliers to 31.0–77.7% with 10% outliers (see **Table 2**). For RMSEA and CFI, the impact of N was smaller: when $N = 500$, median RMSEA increased from 0.000 ($SD = 0.012$) to 0.050 ($SD = 0.035$); median CFI decreased from 1.00 ($SD = 0.004$) to 0.972 ($SD = 0.066$). Simiar trends were observed for $N = 100$ and $N = 200$.

With ML-*t*, LRT, RMSEA, and CFI were relatively stable with increasing proportion of outliers. When $N = 100$, LRT were similar with no outliers, median = 23.15 ($SD = 7.26$), and with 10% outliers, 24.53 ($SD = 7.51$); when $N = 500$, median LRT increased slightly from 22.56 ($SD = 6.81$) with no outliers to 28.71 ($SD = 8.63$) with 10% outliers. Empirical Type I error rates were inflated from 5.2 to 7.2% with no outliers to 10.5–22.7% with 10% outliers. For RMSEA and CFI, when $N = 500$, median RMSEA increased from 0.000 ($SD = 0.012$) to 0.022 ($SD = 0.015$); median CFI decreased from 1.00 ($SD = 0.004$) to 0.994 ($SD = 0.008$). Simiar trends were observed for $N = 100$ and $N = 200$.

Outliers with misspecified model

Figure 3 showed the boxplots of sample values of LRT, RMSEA, and CFI for conditions with a misspecified model and the presence of outliers. In general, the patterns were similar to those observed with a correctly specified model, except that, predictably, the fit was worse on all conditions. With ML-Normal, when $N = 100$, median LRT increased slightly from 32.45 ($SD = 9.37$) with no outliers to 36.84 ($SD = 20.88$) with 10% outliers; when $N = 500$, median LRT increased from 66.52 ($SD = 15.33$) with no outliers to 86.52 ($SD = 71.38$) with 10% outliers. Empirical power of LRT was inflated from 34.6 to 99.1% with no outliers to 51.5–99.8% with 10% outliers (see **Table 2**). For RMSEA and CFI, when $N = 500$, median RMSEA increased from 0.060 ($SD = 0.011$) to 0.072 ($SD = 0.026$); median CFI decreased from 0.962 ($SD = 0.013$) to 0.938 ($SD = 0.064$). Simiar trends were observed for $N = 100$ and $N = 200$.

With ML-*t*, LRT, RMSEA, and CFI were relatively stable with increasing proportion of outliers and remained close to the population values without data contamination, with medians and

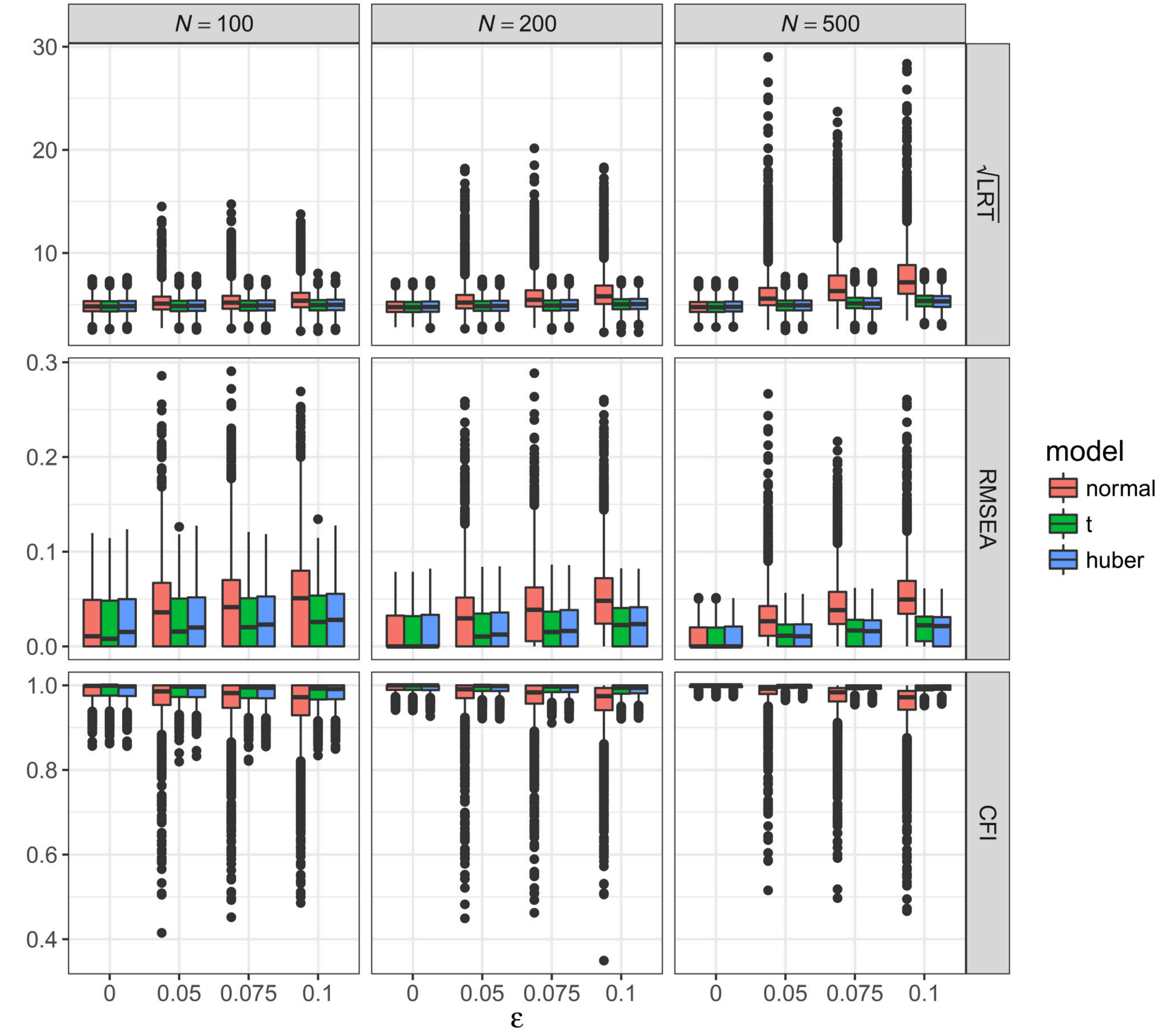


FIGURE 2 | Values of fit indices across simulation conditions with a correctly specified model and the presence of outliers. LRT, likelihood ratio test; ε, proportion of outliers. Square root was taken on the LRT values to reduce its skewness for better graphical presentations.

*SD*s stayed virtually the same regardless of ε. Empirical power was 32.7–99.0% with no outliers and 31.8–98.3% with 10% outliers.

Bad influential observations with correctly specified model

Figure 4 showed the boxplots of sample values of LRT, RMSEA, and CFI for conditions with a correctly specified model and the presence of bad influential observations. Given the nature of such observations, their presence made a bigger impact on LRT, RMSEA, and CFI than outliers did. When $N = 100$, median LRT increased slightly from 23.26 ($SD = 7.28$) with no influential observations to 47.02 ($SD = 41.39$) with 10% influential observations; when $N = 500$, median LRT increased dramatically from 22.73 ($SD = 6.82$) with no influential observations to 137.98 ($SD = 139.53$) with 10% influential observations. Empirical Type I error rates were inflated from 5.4 to 7.3% with no influential observations to 67.8–97.7% with 10% influential observations (see **Table 2**). When $N = 500$, median RMSEA increased from 0.000 ($SD = 0.012$) to 0.100 ($SD = 0.047$), and median CFI decreased from 1.00 ($SD = 0.004$) to 0.915 ($SD = 0.071$). Simiar trends were observed for $N = 100$ and $N = 200$.

With ML-*t*, LRT, RMSEA, and CFI were relatively stable with increasing proportion of influential observations, with medians and *SD*s stayed virtually the same regardless of ε. Empirical Type I error rates were 4.4–7.1% with no influential observations and 4.2–7.2% with 10% influential observations.

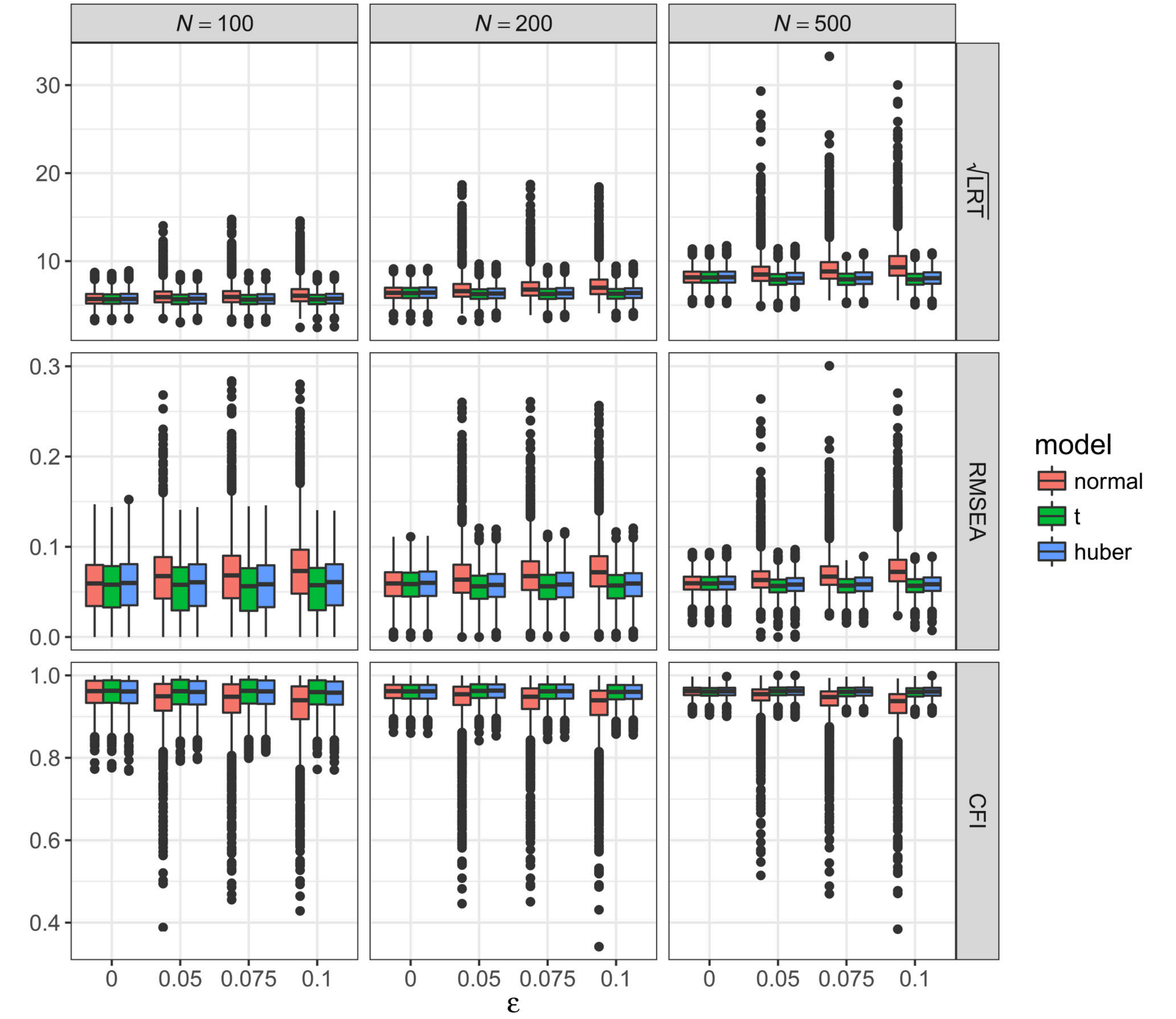


FIGURE 3 | Values of fit indices across simulation conditions with a misspecified model and the presence of outliers. LRT, likelihood ratio test; ε, proportion of outliers. Square root was taken on the LRT values to reduce its skewness for better graphical presentations.

Bad influential observations with misspecified model

Figure 5 showed the boxplots of sample values of LRT, RMSEA, and CFI for conditions with a misspecified model and the presence of bad influential observations. In general, the patterns were similar to those observed with a correctly specified model, except that, predictably, the fit was worse on all conditions. With ML-Normal, when $N = 100$, median LRT increased from 32.45 ($SD = 9.37$) with no outliers to 56.60 ($SD = 43.27$) with 10% influential observations; when $N = 500$, median LRT increased from 66.52 ($SD = 15.33$) with no outliers to 186.28 ($SD = 153.70$) with 10% influential observations. Empirical power was inflated from 34.6 to 99.1% with no outliers to 82.3–100.0% with 10% outliers (see **Table 2**). For RMSEA and CFI, when $N = 500$, median RMSEA increased from 0.060 ($SD = 0.011$) to 0.116 ($SD = 0.043$); median CFI decreased from 0.962 ($SD = 0.013$) to 0.882 ($SD = 0.073$). Simiar trends were observed for $N = 100$ and $N = 200$.

With ML-*t*, LRT, RMSEA, and CFI were relatively stable with increasing proportion of influential observations and remained close to the population values without data contamination, with medians and *SD*s stayed virtually the same regardless of ε. Empirical power was 32.7–99.0% with no outliers and 29.4–97.6% with 10% outliers.

Information Criteria

Figure 6 showed the proportion of replications where AIC, BIC, and SABIC favored ML-*t* over ML-Normal for conditions with

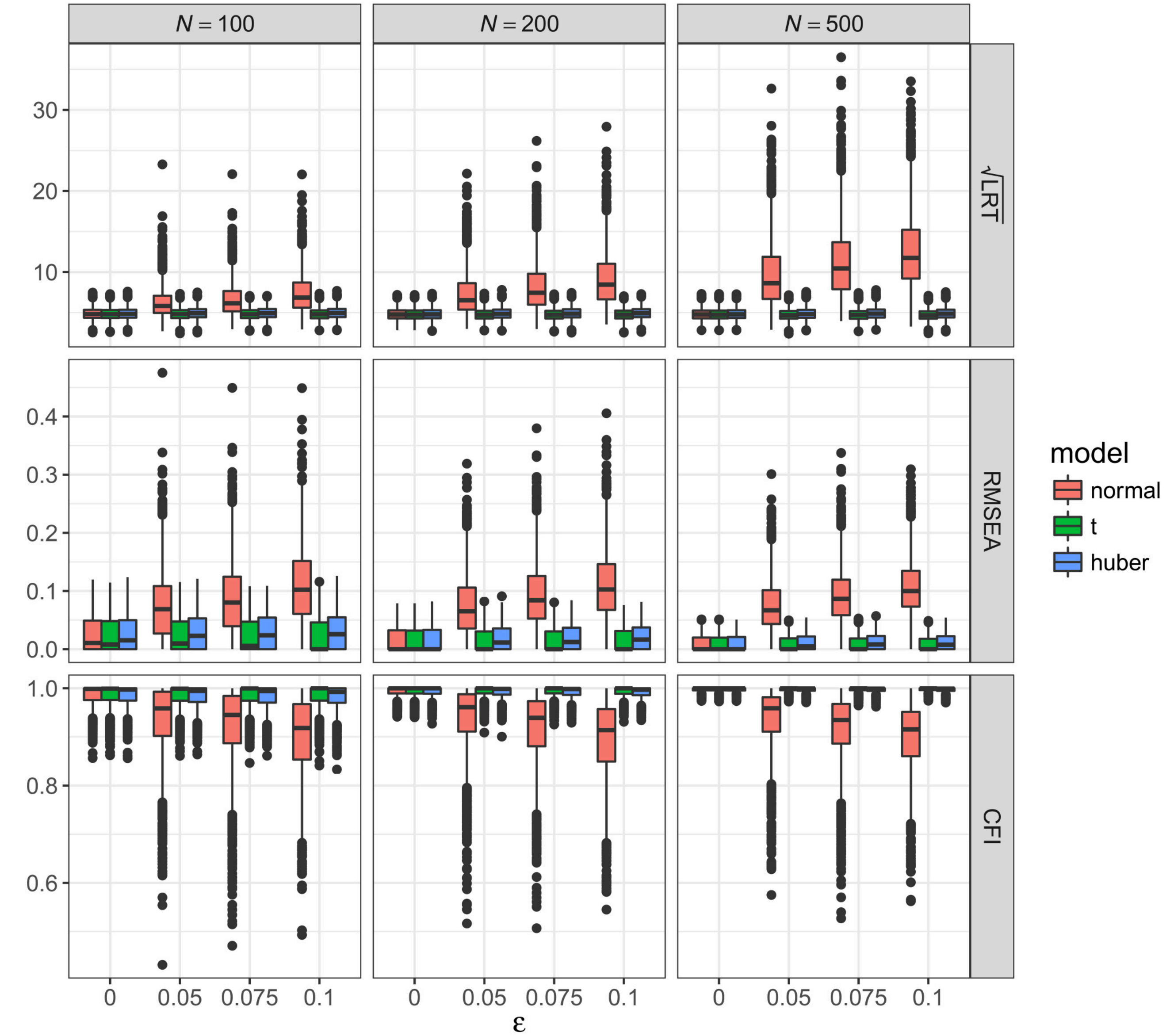


FIGURE 4 | Values of fit indices across simulation conditions with a correctly specified model and the presence of bad influential observations. LRT, likelihood ratio test; ε, proportion of bad influential observations. Square root was taken on the LRT values to reduce its skewness for better graphical presentations.

a correctly specified model, and the results were essentially identical for conditions with a misspecified model. Under a correctly specified model, with no outliers or influential observations, in only 3.6–4.7% of the replications ML-*t* was preferred over ML-Normal by AIC, 0.3–0.9% by BIC, and 2.4–5.4% by SABIC; with 10% outliers, ML-*t* was preferred more often with increasing proportion of outliers and with larger N, with AIC, BIC, and SABIC favoring ML-*t* in 97.5, 94.1, and 96.7% of the replications when $N = 500$. Similarly, with influential observations, AIC, BIC, and SABIC preferred ML-*t* in 76.5, 69.2, and 79.1% of the replications when $\varepsilon = 0.05$ and $N = 100$ and well above 90% for all conditions with either $\varepsilon = 0.10$ or $N \geq 200$.

DISCUSSION

Although the impact of and ways to handle outliers and influential observations have received much attention in regression literature, relatively less discussions on those issues were found in the context of SEM. As pointed out in Yuan and Zhong (2013), unlike general statistics software where diagnostic tools for outliers and influential observations are common, such tools are rarely accessible for SEM software, partly because of the complexity of SEM modeling. Whereas robust SEM using Huber-type weights has been developed and shown to perform well, and the `rsem` package is freely available in R, many researchers are more familiar with other commonly used SEM software packages such as Mplus, and so it is important to have comparable tools for

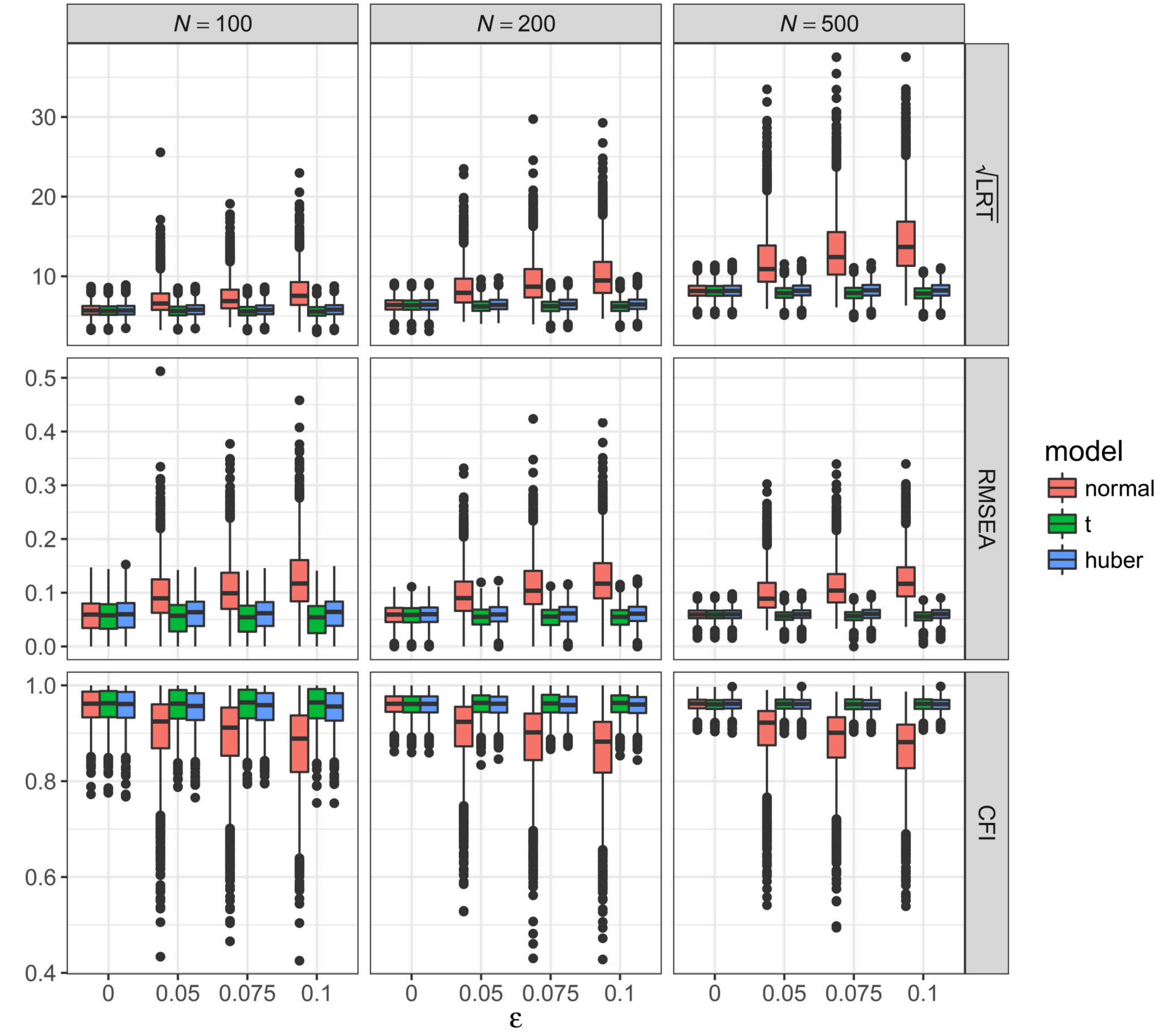


FIGURE 5 | Values of fit indices across simulation conditions with a misspecified model and the presence of bad influential observations. LRT, likelihood ratio test; ε, proportion of bad influential observations. Square root was taken on the LRT values to reduce its skewness for better graphical presentations.

handling outliers and influential observations in other software. With the *t*-based model recently added to Mplus, this study brings attention to this easy-to-use strategy to clarify whether suboptimal model fit is due to global misfit or just a small proportion of extreme cases.

Our simulation results showed that outliers and influential observations could hurt model convergence and dramatically make model fit appear worse for both correctly specified and misspecified SEM models with the usual ML estimation assuming normality. For example, with 25 outliers in a sample of 500 observations, the empirical Type I error rate for LRT was inflated to 0.40 from the nominal level of 0.05, and it was inflated to 0.85 with 25 bad influential observations. Both RMSEA and CFI were more likely to indicate worse model fit in the presence of outliers and influential observations, as predicted in Yuan and Zhong (2013). As it was not common that applied researchers check for outliers and influential observations when conducting SEM (Aguinis et al., 2013), such extreme values may make researchers reject models with adequate fit or consider alternative models that improve overall model fit mainly because of those few observations.

On the other hand, the multivariate-*t* model as well as the two-stage robust method were more robust to data contamination, producing fit indices that were closer to what could have been obtained without those extreme values. First, when sample size was small ($N = 100$) ML-*t* may have some convergence problems

TABLE 2 | Rejection rates of the likelihood ratio test across conditions.

				Outliers			Bad Influential Observations		
Model	***N***	**Estimation**	**$\varepsilon = 0$**	**$\varepsilon = 0.05$**	**$\varepsilon = 0.075$**	**$\varepsilon = 0.10$**	**$\varepsilon = 0.05$**	**$\varepsilon = 0.075$**	**$\varepsilon = 0.10$**
Correct	100	ML-Normal	0.07	0.20	0.23	0.31	0.47	0.55	0.68
		ML-*t*	0.07	0.09	0.08	0.10	0.07	0.07	0.07
		TSR	0.08	0.09	0.09	0.11	0.09	0.09	0.11
	200	ML-Normal	0.06	0.25	0.37	0.46	0.62	0.76	0.84
		ML-*t*	0.06	0.08	0.11	0.12	0.05	0.05	0.05
		TSR	0.07	0.08	0.11	0.12	0.08	0.08	0.09
	500	ML-Normal	0.05	0.40	0.61	0.78	0.85	0.94	0.98
		ML-*t*	0.05	0.09	0.17	0.23	0.05	0.06	0.05
		TSR	0.06	0.09	0.16	0.22	0.08	0.09	0.09
Misspecified	100	ML-Normal	0.35	0.44	0.46	0.52	0.67	0.74	0.82
		ML-*t*	0.33	0.32	0.30	0.32	0.31	0.30	0.29
		TSR	0.34	0.35	0.33	0.35	0.39	0.37	0.40
	200	ML-Normal	0.66	0.72	0.77	0.82	0.90	0.94	0.97
		ML-*t*	0.66	0.60	0.60	0.63	0.59	0.59	0.58
		TSR	0.67	0.64	0.64	0.65	0.67	0.69	0.69
	500	ML-Normal	0.99	0.99	1.00	1.00	1.00	1.00	1.00
		ML-*t*	0.99	0.98	0.98	0.98	0.98	0.97	0.98
		TSR	0.99	0.98	0.99	0.99	0.99	0.99	0.99

Note. ε, Proportion of outliers and bad influential observations; ML, maximum likelihood estimation; TSR, Two-stage robust methods with Huber-type weights downweighing 10% of observations.

in 5–7% of the replications with no model misspecifications and in 9–12% of the replications with misspecifications; however, with $N = 200$ or above the use of ML-t had improved convergence rates over ML-Normal. Second, although to a much less degree, with ML-t and TSR, LRT still increased with increasing proportion of outliers, and empirical Type I error rates increased to 0.10 and 0.11 for $N = 100$ and 0.23 and 0.22 for $N = 500$ with 10% of outliers. Although this is certainly not ideal, LRT under ML-t or TSR still performs much better than under ML-Normal. Future studies can focus on how to obtain adjusted test statistics for ML-t. Note, however, when the model was misspecified, or when the extreme values were bad influential observations, LRT, RMSEA, and CFI were all similar regardless of proportions of data contamination, and the values under ML-t were slightly closer to the population values than those under TSR.

Third, information criteria was effective in picking ML-Normal when no outliers or bad influential observations were present in the data, and in picking ML-t when extreme values were present, with better accuracy when sample size increased. Under our simulation conditions, we found AIC and SABIC showed higher sensitivity than BIC. Therefore, when researchers are uncertain whether data contamination could be a problem, an effective way in determining whether to use ML-Normal or ML-t is to choose one that gives smaller AIC and SABIC.

It is generally recommended to use multivariate-t-based SEM and other robust SEM methods, rather than directly deleting outliers and influential observations, as the complexity of SEM makes it more challenging to use general techniques such as Mahalanobis distance and Cook's distance to identify outliers and influential observations (e.g., Flora et al., 2012; Sterba and Pek, 2012). Although these methods provided parameter estimates and fit indices that are insensitive to the influence of outliers, they do not replace the need for careful data screening work. As suggested by Aguinis et al. (2013), one should always identify in the data if there are any extreme cases due to correctible errors, and correct them accordingly. Failure to do so may lead to loss of valuable information. Also, after any such errors are corrected, the use of robust SEM is justified only when the outliers and influential observations are regarded as coming from a different data generating process than the majority of the data, and the goal of inference is to estimate a model that is representative of most of the data. Sometimes outliers and influential observations can be of interest in their own rights, and they can lead to important research findings (Aguinis et al., 2013; O'Connell et al., 2015). Also, a non-trivial proportion of such cases may indicate unmodeled heterogeneity, where the use of mixture models may be more appropriate. Recent methodological work has provided accessible tools to identify outliers and influential observations for SEM (Pek and MacCallum, 2011; Sterba and Pek, 2012), which we recommend to be used in combination with robust SEM methods.

Despite the contributions of the study, there are several limitations that call for future studies. First, as a first step to evaluate the multivariate-t-based SEM, we chose to first study the performance of fit indices under such a model. An obvious next step is to make sure that the parameter estimates are sensible with the multivariate-t-based SEM, which appeared to be robust

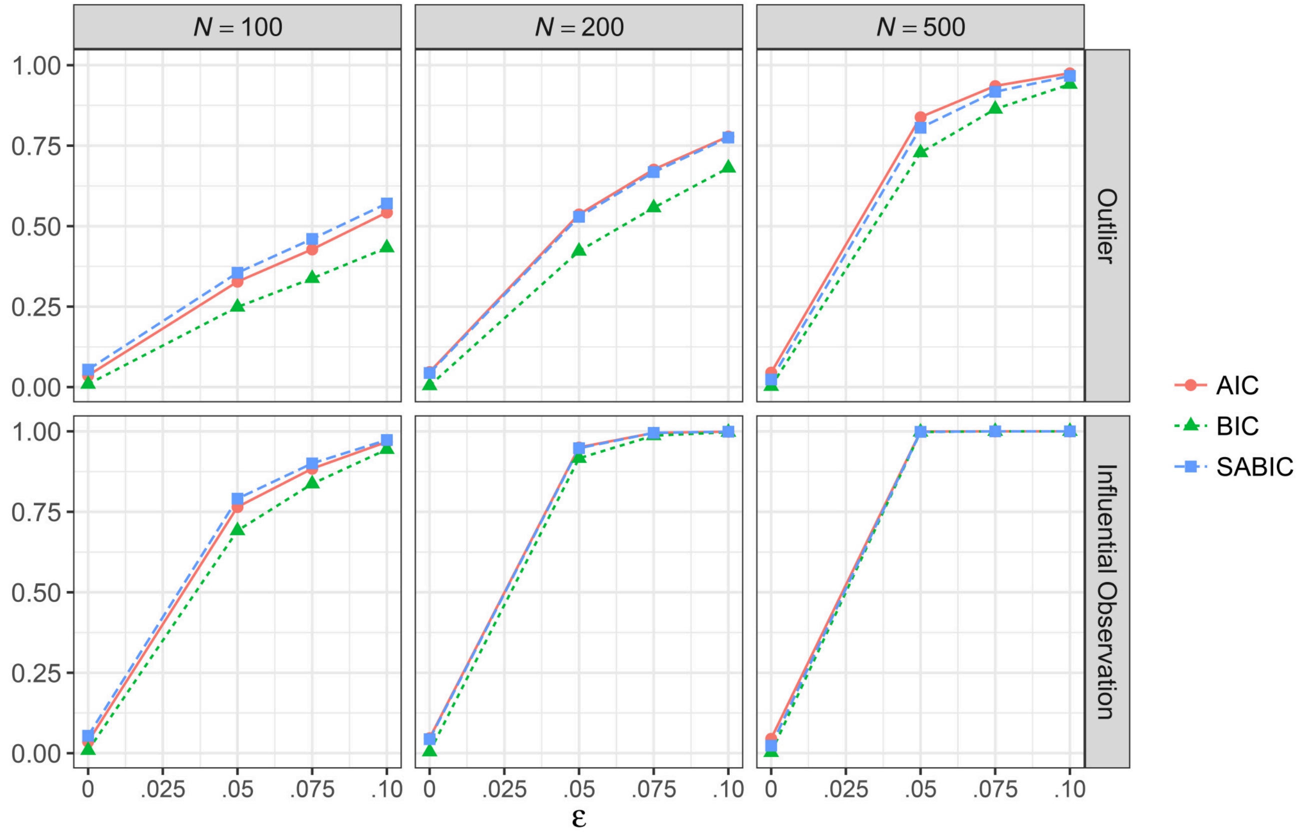


FIGURE 6 | Proportion of replications where the multivariate *t* model has smaller information criteria than the multivariate normal model across simulation conditions with a correctly specified model. AIC, Akaike information criteria; BIC, Bayesian information criteria; SABIC, sample-size adjusted information criteria; ε, proportion of outliers.

based on the real data example and the results in Yuan and Bentler (1998b) using weights corresponding to a multivariate *t* distribution with one degree of freedom. Second, as pointed out in Yuan et al. (2004), the use of multivariate-*t*-based SEM might not be as efficient as the use of Huber-type weights under some conditions, and future studies may compare the performance of various robust methods in simulated and real data. At this stage, we found that the use of the multivariate-*t*-based SEM is accessible to researchers without the need to choose a tuning parameter, allows conventional interpretations of information criteria, and can be easily integrated into more complex SEM models.

Third, it should be emphasized again that, in the current study, we only focused on situations where a small proportion of data is contaminated, whereas the majority of the data still satisfies the normality assumption. Although the resulting data also had skewness and kurtosis deviated from those of a normal distribution, common SEM estimation methods that are robust to non-normality may not work well in the presence of data contamination. Whereas corrections for non-normality such as the Satorra-Bentler procedure relies on sandwich estimator and higher-order moments of the sample data, ML-*t* as implemented in Mplus uses maximum likelihood with the expectation-maximization algorithm to estimate the model parameters, including degrees of freedom. To examine our speculations, we re-analyze the simulated data using the Satorra-Bentler correction procedure (`ESTIMATOR=MLM` in Mplus), and found the resulting fit indices to still be sensitive to data contamination, although not to the extent as ML-Normal. For example, with $N = 500$, 10% outliers, and a correctly specified model, median RMSEA = 0.041 with Satorra-Bentler, commpared to 0.050 for ML-Normal and 0.022 for ML-*t*. Therefore, researchers should distinguish between robustness against data contamination, which can be handled with ML-*t* or Huber-type weights, and robustness against non-normality, which can be alleviated by the Satorra-Bentler correction or weighted least squares estimator.

Fourth, our simulations did not cover sample sizes smaller than 100. We had performed additional simulations with $N = 50$ and found that TSR had very low convergence rates (less than 5%) and ML-*t* had convergence around 71–76%, and information criteria preferred ML-*t* only 20% of the replications in the presence of 10% outliers observations. Therefore, we recommend using ML-*t* or the two-stage robust methods only with a sample size of at least 100. Finally, in this study we only evaluated fit indices with a factor model, and consider only misspecifications in the form of missing one cross-loading. As the impact of outliers and influential observations on fit indices may vary depending on types of SEM models, model complexity, and parameter values, future studies can expand on the simulation conditions to provide more complete information on this underresearched area.

AUTHOR CONTRIBUTIONS

ML designed the simulation and drafted the manuscript. JZ helped conducting the simulation, provided recommendations on the draft, and proofread.

REFERENCES

Aguinis, H., Gottfredson, R. K., and Joo, H. (2013). Best-practice recommendations for defining, identifying, and handling outliers. *Org. Res. Methods* 16, 270–301. doi: 10.1177/1094428112470848

Asparouhov, T., and Muthén, B. (2015). Structural equation models and mixture models with continuous non-normal skewed distributions. *Struc. Eq. Model. A Multidiscip. J.* 23, 1–19. doi: 10.1080/10705511.2014.947375

Bentler, P. M. (1983). Some contributions to efficient statistics in structural models: specification and estimation of moment structures. *Psychometrika* 48, 493–517. doi: 10.1007/BF02293875

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *Br. J. Math. Stati. Psychol.* 37, 62–83. doi: 10.1111/j.2044-8317.1984.tb00789.x

Flora, D. B., LaBrish, C., and Chalmers, R. P. (2012). Old and new ideas for data screening and assumption testing for exploratory and confirmatory factor analysis. *Front. Psychol.* 3:55. doi: 10.3389/fpsyg.2012.00055

Gelman, A., and Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University.

Holzinger, K. J., and Swineford, F. (1939). *A Study in Factor Analysis: The Stability of a Bi-factor Solution (Supplementary Educational Monograph No. 48)*. Chicago, IL: University of Chicago.

Jackson, D. L., Gillaspy, J. Arthur, J., and Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: an overview and some recommendations. *Psychol. Methods* 14, 6–23. doi: 10.1037/a0014694

Kline, R. B. (2011). *Principles and Practice of Structural Equation Modeling, 3 Edn*. New York, NY: Guilford.

Muthén, L. K., and Muthén, B. O. (1998–2015). *Mplus User's Guide, 7 Edn*. Los Angeles, CA: Muthén & Muthén.

O'Connell, A. A., Yeomans-Moldanado, G., and McCoach, D. B. (2015). "Residual diagnostics and model assessment in a multilevel framework: recommendations toward best practice," in *Advances in Multilevel Modeling for Educational Research: Addressing Practical Issues Found in Real-World Applications*, eds J. R. Harring, L. M. Stapleton, and S. N. Beretvas (Charlotte, NC: Information Age), 97–135.

Pek, J., and MacCallum, R. C. (2011). Sensitivity analysis in structural equation models: cases and their influence. *Multiv. Behav. Res.* 46, 202–228. doi: 10.1080/00273171.2011.561068

Pinheiro, J. C., Liu, C., and Wu, Y. N. (2001). Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate *t* distribution. *J. Comput. Graph. Stat.* 10, 249–276. doi: 10.1198/10618600152628059

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Satorra, A., and Bentler, P. M. (1994). "Corrections to test statistics and standard errors in covariance structure analysis," in *Latent Variables Analysis: Applications to Developmental Research*, eds A. von Eye and C. C. Clogg (Thousand Oaks, CA: Sage), 339–419.

Sterba, S. K., and Pek, J. (2012). Individual influence on model selection. *Psychol. Methods* 17:582. doi: 10.1037/a0029253

Yuan, K.-H., and Bentler, P. M. (1998a). Robust mean and covariance structure analysis. *Br. J. Mat. Stat. Psychol.* 51, 63–88. doi: 10.1111/j.2044-8317.1998.tb00667.x

Yuan, K.-H., and Bentler, P. M. (1998b). Structural equation modeling with robust covariances. *Sociol. Methodol.* 28, 363–396. doi: 10.1111/0081-1750.00052

Yuan, K.-H., and Bentler, P. M. (2000). Robust mean and covariance structure analysis through iteratively reweighted least squares. *Psychometrika* 65, 43–58. doi: 10.1007/BF02294185

Yuan, K.-H., and Bentler, P. M. (2001). Effect of outliers on estimators and tests in covariance structure analysis. *Br. J. Math. Stat. Psychol.* 54, 161–175. doi: 10.1348/000711001159366

Yuan, K.-H., Bentler, P. M., and Chan, W. (2004). Structural equation modeling with heavy tailed distributions. *Psychometrika* 69, 421–436. doi: 10.1007/BF02295644

Yuan, K. H., Chan, W., and Bentler, P. M. (2000). Robust transformation with applications to structural equation modelling. *Br. J. Math. Stat. Psychol.* 53, 31–50. doi: 10.1348/000711000159169

Yuan, K.-H., and Hayashi, K. (2010). Fitting data to model: structural equation modeling diagnosis using two scatter plots. *Psychol. Methods* 15, 335–351. doi: 10.1037/a0020140

Yuan, K.-H., and Zhang, Z. (2012). Structural equation modeling diagnostics using R package semdiag and EQS. *Struc. Eq. Modeli. A Multidisc. J.* 19, 683–702. doi: 10.1080/10705511.2012.713282

Yuan, K.-H., and Zhang, Z. (2015). *rsem: Robust Structural Equation Modeling with Missing Data and Auxiliary Variables*. R package version 0.4.6. Available online at: https://CRAN.R-project.org/package=rsem

Yuan, K.-H., and Zhong, X. (2008). Outliers, leverage observations, and influential cases in factor analysis: Using robust procedures to minimize their effect. *Sociol. Methodol.* 38, 329–368. doi: 10.1111/j.1467-9531.2008.00198.x

Yuan, K.-H., and Zhong, X. (2013). Robustness of fit indices to outliers and leverage observations in structural equation modeling. *Psychol. Methods* 18, 121–136. doi: 10.1037/a0031604

Zellner, A. (1976). Bayesian and non-Bayesian analysis of the regression model with multivariate Student-*t* error terms. *J. Am. Stat. Assoc.* 71, 400–405. doi: 10.2307/2285322

Zhong, X., and Yuan, K.-H. (2011). Bias and efficiency in structural equation modeling: maximum likelihood versus robust methods. *Multiv. Behav. Res.* 46, 229–265. doi: 10.1080/00273171.2011.558736

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

METHODS
published: 27 December 2016
doi: 10.3389/fpsyg.2016.01973

Practical Person-Fit Assessment with the Linear FA Model: New Developments and a Comparative Study

Pere J. Ferrando *, *Andreu Vigil-Colet and Urbano Lorenzo-Seva*

Research Center for Behavior Assessment, Department of Psychology, Universitat Rovira I Virgili, Tarragona, Spain

Linear factor analysis (FA) is, possibly, the most widely used model in psychometric applications based on graded-response or more continuous items. However, in these applications consistency at the individual level (person fit) is virtually never assessed. The aim of the present study is to propose a simple and workable approach to routinely assess person fit in FA-based studies. To do so, we first consider five potentially appropriate indices, of which one is a new proposal and the other is a modification of an existing index. Next, the effectiveness of these indices is assessed by using (a) a thorough simulation study that attempts to mimic realistic conditions, and (b) an illustrative example based on real data. Results suggest that the mean-squared *lico* index and the personal correlation work well in conjunction and can function effectively for detecting different types of inconsistency. Finally future directions and lines of research are discussed.

Keywords: person-fit statistics, linear factor analysis, mean-squared person-fit indices, personal correlation, outliers detection

OPEN ACCESS

Edited by:
Mike W.-L. Cheung,
National University of Singapore (NUS), Singapore

Reviewed by:
Daniel Oberski,
Tilburg University, Netherlands
Rob Meijer,
University of Groningen, Netherlands

Correspondence:
Pere J. Ferrando
perejoan.ferrando@urv.cat

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *07 November 2016*
Accepted: *05 December 2016*
Published: *27 December 2016*

Citation:
Ferrando PJ, Vigil-Colet A and Lorenzo-Seva U (2016) Practical Person-Fit Assessment with the Linear FA Model: New Developments and a Comparative Study.
Front. Psychol. 7:1973.
doi: 10.3389/fpsyg.2016.01973

INTRODUCTION

When used for item analysis and individual scoring purposes, the standard factor-analysis (FA) model can be viewed as a linear item response theory (IRT) model intended for continuous scores (e.g., Ferrando, 2009). In practice it is generally used with discrete item scores and in these cases it can be only approximately correct. However, for graded-response or more continuous item formats, the linear FA approximation has proved to be reasonably good in many conditions that can be found in practice (Hofstee et al., 1998; Ferrando, 2009; Rhemtulla et al., 2012; Culpepper, 2013; Ferrando and Lorenzo-Seva, 2013). Furthermore, in comparison to the theoretically more appropriate nonlinear models, linear FA has the non-negligible advantages of simplicity, and robustness (e.g., Briggs and MacCallum, 2003; Ferrando and Lorenzo-Seva, 2013).

The appropriateness of the FA model is usually assessed by conducting an overall goodness-of-fit investigation based on the entire dataset (e.g., Reise and Widaman, 1999). Model-data fit, however, can also be assessed at the individual-level, by considering the responses of each individual across the set of test items. This level of assessment, which is usually known as "person fit," is almost always neglected in psychometric FA applications, and is the topic of the present article.

Person-fit analysis refers to a variety of indices and procedures aimed at assessing the fit of each individual score pattern to the psychometric model fitted to the data (see e.g., Meijer et al., 2015). This type of assessment is generally sequential (e.g., Rupp, 2013; Conijn et al., 2015; Ferrando, 2015; Meijer et al., 2015), and the simplest schema is two-stage. In the first stage, a global or practical index is used to flag potentially inconsistent respondents without specifying the kind of inconsistency. In the second stage, a more specific analysis is carried out in order to ascertain the sources and effects of misfit in those patterns that are flagged as potentially inconsistent. Here we shall only consider practical indices to be used in the first stage.

Person-fit assessment is important for various reasons (see e.g., Reise and Widaman, 1999; Meijer et al., 2015) but mainly for a practical validity reason: if a response pattern is not well explained by the model, there is no guarantee that the score assigned to this pattern will adequately reflect the "true" trait level of the individual. So, this score cannot be validly interpreted. This compelling reason requires individual response patterns to be routinely checked so that invalid test scores can be detected (e.g., International Test Commission, 2014; Tendeiro and Meijer, 2014). In IRT applications, however, this recommendation is far from common practice (Meijer et al., 2015), and practical person-fit indices appear to be used routinely only in Rasch-based applications, possibly because they have been implemented and provided as standard output in these computer programs ever since they have been available (Wright et al., 1979; Smith, 1986).

The main contention of this article is that routine FA-based person fit assessment will only become (hopefully) common practice if (a) a clear proposal based on simple, effective and easily interpretable practical indices is made, and (b) this proposal is implemented in a free, user-friendly program that is easily available.

In principle, the procedures considered here could be (a) applied to both unidimensional and multidimensional solutions, and (b) used in both typical-response (personality and attitude) and ability measurement (e.g., Clark, 2010). For the moment, however, we shall focus only on unidimensional solutions intended for typical-response items. As for the first restriction, the unidimensional model is the simplest and the most univocally interpretable, and, therefore, is expected to lead to clearer results regarding person-fit assessments (e.g., Conijn et al., 2014). As for the second, most of the existing measures based on graded or more continuous items are typical-response (e.g., Ferrando, 2009).

REVIEW OF BASIC FA RESULTS

Consider a questionnaire made up of n items with (approximately) continuous responses that intends to measure a single trait or common factor θ. For a person i who responds to an item j, the linear FA model is:

$$X_{ij} = \mu_j + \lambda_j\theta_i + \varepsilon_{ij} \tag{1}$$

where: X_{ij} is the observed item score, μ_j is the item intercept, λ_j the item loading, ϵ_{ij} the measurement error, and θ is scaled in a z-score metric (mean 0 and variance 1). For fixed θ, the item scores are distributed independently (local independence), and the conditional distribution is assumed to be normal, with mean and variance given by

$$\hat{X}_{ij} = E(X_j \mid \theta_i) = \mu_j + \lambda_j\theta_i \quad ; \quad Var(X_j|\theta) = \sigma^2_{\varepsilon j} \tag{2}$$

If the item and person parameters in Equations (1) and (2) are known, it then follows that the standardized residual:

$$z_{ij} = \left(\frac{X_{ij} - \hat{X}_{ij}}{\sigma_{\varepsilon j}}\right) \tag{3}$$

is a value drawn at random from the standard normal distribution. By the local independence principle, it then follows that the sum:

$$S_i = \sum_j^n z^2{}_{ij} \tag{4}$$

is distributed as χ^2 with n degrees of freedom. So, $E(S_i) = n$, and $\mathrm{Var}(S_i) = 2n$.

In most practical applications, neither the structural parameters (μ_j,λ_j, and $\sigma^2_{\epsilon j}$) nor the "true" trait levels θ_i are known, and they have to be estimated. We shall assume here that model (1) is fitted using a standard two-stage procedure (McDonald, 1982). In the first stage (item calibration), the structural (item) parameters are estimated. In the second stage (scoring), the item estimates are taken as fixed and known, and used to obtain trait estimates or factor scores for each individual. We shall further assume that the individual trait estimates are maximum likelihood (ML) estimates, given by

$$\hat{\theta}_i(ML) = \frac{\sum_j^n \frac{\lambda_j(X_{ij} - \mu_j)}{\sigma_{\varepsilon j}{}^2}}{\sum_j^n \frac{\lambda_j{}^2}{\sigma_{\varepsilon j}{}^2}} \tag{5}$$

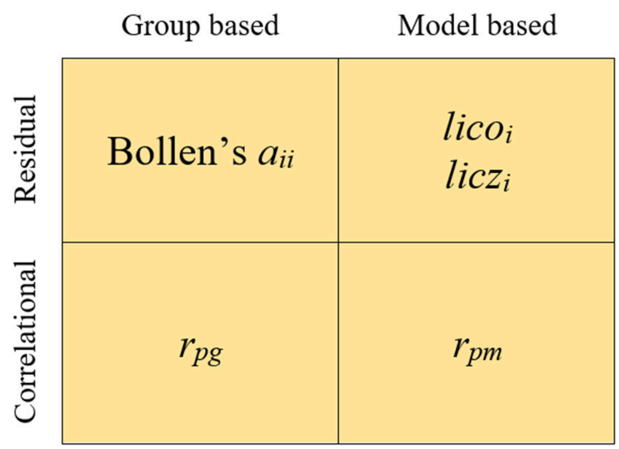
In FA terminology, the estimates in Equation (5) are known as Bartlett's weighted least squares factor scores (e.g., McDonald, 1982).

OVERVIEW OF THE SELECTED INDICES AND RATIONALE

The indices we shall consider in the study fall into four different categories which arise when two different criteria are combined. The resulting categories and indices are summarized in **Figure 1**.

The first criterion distinguishes between model-based (MB) or parametric vs. model-free or group-based (GB) indices. In MB indices, the information provided by the parameter estimates of the model is used to assess person fit. In the case of FA this information refers to (a) the item parameter estimates and (b) the individual trait level estimate or factor score. In contrast, the GB indices use only the information provided by the responses of the group of individuals which is assessed. So, the fit of the response pattern is assessed with respect to the majority of response patterns in the group (e.g., Tendeiro and Meijer, 2014).

Because MB indices use more information than GB indices they should be more powerful. In simulation studies, however, it is not unusual for GB indices to outperform their theoretically



	Group based	Model based
Residual	Bollen's a_{ii}	$lico_i$ $licz_i$
Correlational	r_{pg}	r_{pm}

FIGURE 1 | Indices used in the study.

superior counterparts (Karabatsos, 2003; Tendeiro and Meijer, 2014; Meijer et al., 2015). This result does have some plausible explanations. First, the presence of some inconsistent respondents might distort the structural (i.e., item) estimates (Nering, 1997). Second, the same response vector that is used to obtain the trait estimate is then used to assess person misfit. So, if the response vector does not fit, the inconsistency is likely to bias the trait estimate and this bias, in turn, will distort the MB person-fit value in the direction of making the response vector appear less inconsistent than it really is (Karabatsos, 2003; Armstrong et al., 2007). The source of this second problem is, indeed, that the true trait levels are unknown, so estimates (ML in our case) are used in their place. In general, the closer the estimates are to the true values, the more effective the MB indices will be at detecting inconsistencies (Reise, 1995). However, to one extent or another, trait estimates are unreliable and indeterminate (i.e., the problem of factor indeterminacy, see Guttman, 1955), and the more unreliable and indeterminate they are, the less effective the MB indices based on them are expected to be.

The second criterion in **Figure 1** distinguishes between residual vs. correlational indices. Residual indices are generally mean-squared measures that assess the discrepancies between the observed and the expected (from the model estimates or from the group responses) response vectors. Correlational indices are based on the product-moment correlation between the observed-expected vectors.

The relations between residual and correlational indices can be discussed by using some basic concepts from profile analysis. The residual indices that we shall consider here are D^2–type indices (Cronbach and Gleser, 1953), based on the squared distance between the observed and the expected vectors. So, they simultaneously consider differences in elevation (score means), scatter or dispersion (score standard deviations) and shape (mainly rank ordering agreement between observed and expected scores). In contrast, correlational indices are only affected by differences in shape. So, in principle residual indices should be more powerful than correlational indices because they use more information from the data. Again, however, the simpler correlational indices have performed surprisingly well in some simulation studies (Rudner, 1983).

RESIDUAL-BASED INDICES

GB Indices

In the more general field of outlier detection, Bollen (1987) proposed a model-free residual statistic which is, essentially, a scaled Mahalanobis distance based on an unstructured covariance matrix (e.g., Yuan et al., 2004). Denote by **Z**, of dimension $N \times n$, the matrix containing all the person × item scores written as deviations from the variable means. Next, define the $N \times N$ **A** matrix as

$$\mathbf{A} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}' \qquad (6)$$

The elements a_{ii} in the main diagonal of **A** are Bollen's person-fit indices for the *i* individual. These elements measure the distance of the response vector of individual *i* from the means for all of the items. They tend to flag as potentially inconsistent those cases that sit far away from the center of the data. In terms of interpretation they have two interesting properties: first, they are scaled to provide values in the range 0–1. Second, their average value is *n/N* and this is a reference for judging the magnitude of a_{ii}. The main shortcoming is that an individual with an extreme trait level that responds consistently with the FA model may be flagged as potentially inconsistent with this index.

MB Indices

Several indices have been proposed in this category (Bollen and Arminger, 1991; Yuan et al., 2004). Here, we shall consider an index proposed by Ferrando (2007) denoted here as *lco*. It is the sum of the squared residuals in Equation (4) evaluated by using the ML trait estimate in Equation (5) instead of the unknown "true" trait level.

$$lco_i = \sum_{j}^{n} \frac{(X_{ij} - \mu_j - \lambda_j \hat{\theta}_i(ML))^2}{\sigma^2{}_{\varepsilon j}}. \qquad (7)$$

Because a minimum-chi-square trait estimate is used as a substitute for θ, it follows that, if the model is correct and under the null hypothesis that all the respondents are consistent, the distribution of *lco* is expected to be χ^2 with *n-1* degrees of freedom. So, the expected value of *lco* is *n-1* and its variance is *2(n-1)*. Conceptually *lco* measures discrepancies between an individual pattern of observed scores and the pattern which would be expected from the FA model given the trait estimate for this individual. So, large *lco* values indicate non-fitting response patterns.

Our real-data applications based on *lco* suggest that the index is of practical interest, but they have also revealed a problem of over-sensitivity to unexpected responses in items of good quality (i.e., with a small residual variance). This result can be anticipated by inspecting Equation (7) and is well documented in Rasch analysis, in which discrepancy indices of the form Equation (7) are labeled as "outfit" statistics [meaning outlier-sensitive fit (e.g., Wright and Masters, 1982; Smith et al., 1998)].

In Rasch-based measurement, weighted discrepancy indices labeled as "infit" statistics, have been proposed to counteract the over-sensitivity problem discussed above (Wright and Masters, 1982; Smith et al., 1998). In the same spirit, we propose here a new FA-based weighted statistic which is defined as

$$lico_i = \left(\frac{n}{n-1}\right) \frac{\sum_j^n (X_{ij} - \mu_j - \lambda_j \hat{\theta}_i(ML))^2}{\sum_j^n \sigma^2{}_{\varepsilon j}}. \tag{8}$$

To derive the mean and variance of *lico*, consider first the simple case in which the trait levels are known. In this case, Equation (8) could be written as (see Equations 2, 3):

$$lico_i = \left(\frac{n}{n-1}\right) \frac{\sum_j^n \sigma^2{}_{\varepsilon j} z^2{}_{ij}}{\sum_j^n \sigma^2{}_{\varepsilon j}} = \sum_j^n w_j z^2{}_{ij}, \tag{9}$$

i.e., a linear combination of independent χ^2 variables (z_j^2) each of which has one degree of freedom. So, $E(z_j^2) = 1$ and $Var(z_j^2)$ = 2. By considering next the loss of one degree of freedom when $\hat{\theta}_i(ML)$ is used instead of the unknown θ_i, the mean and variance of *lico* are found to be

$$\begin{aligned} E(lico_i) &= 1 \\ Var(lico_i) &= \left(\frac{n}{n-1}\right) \frac{2\sum_j^n \sigma^4{}_{\varepsilon j}}{\left(\sum_j^n \sigma^2{}_{\varepsilon j}\right)^2}. \end{aligned} \tag{10}$$

Overall, *lico* is a weighted mean-squared statistic which has unit expectation under the null hypothesis of consistency. As in the case of *lco*, large values (in this case larger than the unit reference value) suggest inconsistency. As for cutoff values, in Rasch measurement conventional values of about 1.3–1.5 are generally used for judging potential inconsistency based on this type of statistics (e.g., Wright and Linacre, 1994). However, Equation (9) shows that the expected variance of *lico* (and, therefore, its expected range of values) mainly depends on test length. To see this point more clearly, consider that in the case of parallel items, with equal residual variances, the variance term in (9) reduces to *2/(n-1)*, which is indeed the variance of *lco/(n-1)*. For weighted discrepancy indices based on the Rasch model, Smith et al. (1998) suggested more refined cutoff values that take into account this dependence. They are given by:

$$critical\ value = 1 + \frac{2}{\sqrt{n}}. \tag{11}$$

The appropriateness of this cutoff for the present proposal will be assessed in both the simulation study and the illustrative example.

An alternative possibility in terms of interpretation and cutoff values is to obtain a standardized version of *lico* that can be interpreted as a normal deviate. To do so, we shall consider again the simple case in which the trait levels are known and use the linear-composite expression (9). Jensen and Solomon (1972) found that combinations of this type can be closely approximated to the standard normal by using a Wilson-Hilferty cube-root transformation (Wilson and Hilferty, 1931). Our contention is that this approximation will also be close enough to the normal when ML trait estimates are used instead of unknown true levels. If it is, the new standardized person-fit statistic we propose could be computed as

$$licz_i = ({lico_i}^{1/3} - 1)\left(\frac{3}{\sqrt{Var\left(lico_i\right)}}\right) + \left(\frac{\sqrt{Var\left(lico_i\right)}}{3}\right). \tag{12}$$

In principle, the theoretically-derived *Var(lico$_i$)* is given in Equation (10). However, our preliminary simulation studies suggest that, while the empirical mean value of *lico* is usually quite close to the expected unit value, the empirical variance may be different from the theoretical variance in Equation (10). If it is, the use of the latter is expected to lead to differences between Equation (12) and the reference simulation. To address this problem, we propose to empirically estimate the variance of *lico* by using simulation procedures, and then use this empirical estimate in Equation (12). If it works properly, this combined theoretical-empirical procedure has the advantage that the index can still be interpreted as a normal deviate, with its familiar associated cutoff values that do not depend on test length.

Correlation-Based Indices

Group-Based Indices

Fowler (1954) and Donlon and Fischer (1968) proposed using the correlation between the respondent's response vector and the vector of item sample means as a straightforward person-fit index. This index is usually known as the "personal correlation" and will be denoted here by r_{pg}.

As initially proposed, the personal correlation was only intended for binary responses. Because in this case the value a correlation can have heavily depends on the marginal distribution of the data it is difficult to compare values across persons. Furthermore, there is no standard cutoff value for classifying a respondent as inconsistent on the sole basis of the magnitude of his/her personal correlation. Possibly for these reasons r_{pg} is hardly used nowadays. However, for the approximately continuous item responses considered here, the differential attenuation problem due to marginal differences is considerably minimized. And, regarding the second limitation, r_{pg} might still have an important role as an auxiliary practical index even when there are no simple cutoff values.

Conceptually r_{pg} quantifies the similarity between the item locations for the respondent and the normative item locations obtained from the entire group. In other words, r_{pg} assesses the extent to which the responses of the individual are sensitive to the group-based normative ordering of the items by their extremeness.

Model-Based Indices

We shall propose here a model-based personal correlation index, which we shall denote as r_{pm}, and which is defined as the product-moment correlation between the respondent's response vector ($\mathbf{x}_i$) and the vector of expected item scores ($\hat{x}_i$), whose elements are given by

$$\hat{X}_{ij} = \mu_j + \lambda_j \hat{\theta}_i(ML) \qquad (13)$$

Conceptually r_{pm} measures the similarity (in terms of rank ordering) between the scores obtained by the respondent and the scores that would be expected given the structural FA parameters and his/her trait estimate.

Relations between Residual-Based and Correlation-Based Indices

Within each class, MB and GB, the residual and correlational indices are obtained from the same observed-expected vectors and are algebraically related. The basic relations have been discussed above in terms of profile analysis. In this section we shall further analyse the relations in order to show the complementary role that the residual-based and the correlation-based indices can have in practical assessment. We shall focus the analysis on the relations between r_{pm} and *lico*, which are the most direct ones. The results, however, are still valid in general for both types of index.

By using vector notation and standard covariance algebra, the following result is obtained

$$lico_i = \left(\frac{n^2}{(n-1)\sum_j^n \sigma^2{}_{\varepsilon j}} \right) \Big[(\mathbf{x}_i - \bar{\hat{\mathbf{x}}}_i)^2 + (s(\mathbf{x}_i) - s(\hat{\mathbf{x}}_i))^2 + 2s(\mathbf{x}_i)s(\hat{\mathbf{x}}_i)(1 - r_{pm(i)}) \Big]. \qquad (14)$$

The right hand side of Equation (14) separates the elevation (differences in means), scatter (differences in standard deviations), and differences-in-shape components that are measured by *lico*. If the first two components are kept constant, the relation is indeed negative: the higher r_{pm} is, the lower *lico* Is.

The result (Equation 14) suggests that the effectiveness of the personal correlations and the residual indices will depend on the type of inconsistency. So, if inconsistency mainly affects the rank ordering of the item scores with respect to the group-based normative ordering (r_{pg}) or the model-expected ordering (r_{pm}), then the personal correlations are expected to be more effective than the residual indices. On the other hand, if inconsistency mainly affects the means and variances of the observed-expected vectors, then, residual indices are expected to be more effective. As an example of this second case, consider an extreme respondent who, in everything else, behaves according to the FA model. The expected-observed agreement in terms of rank ordering is perfect in this case. However, the "scatter" and perhaps the "elevation" components differ, because the "high" observed scores are higher than expected while the "low" scores are lower.

We shall finally discuss relations with cutoff values. If the null hypothesis of consistency holds, the expected values of the personal correlations for an individual *i* are found to be:

$$E(r_{pm(i)}) = \sqrt{\frac{\mathrm{var}(\mu_j) + \theta_i^2\, \mathrm{var}(\lambda_j)}{\mathrm{var}(\mu_j) + \theta_i^2\, \mathrm{var}(\lambda_j) + \bar{\sigma}_{\varepsilon j}^2}}$$

and: . (15)

$$E(r_{pg(i)}) = \sqrt{\frac{\mathrm{var}(\mu_j)}{\mathrm{var}(\mu_j) + \theta_i^2\, \mathrm{var}(\lambda_j) + \bar{\sigma}_{\varepsilon j}^2}}$$

For both r_{pg} and r_{pm}, the expected value under the null hypothesis of consistency depends on both the item and the person parameters. So, unlike what occurs with *lico* and *licz*, a simple value cannot be rigorously proposed as a cutoff for r_{pg} and r_{pm}. It is mainly for this reason that we prefer to consider personal correlations as auxiliary indices.

SIMULATION STUDIES

Design and General Conditions

We agree with Rupp (2013) that simulation studies should reflect, as far as possible, the inconsistent behaviors that are found in real life, and we have tried to do this here. Because we are mainly concerned with typical-response measurement (i.e., personality and attitude), we have tried to mimic response mechanisms expected to lead to inconsistent responses in this domain (e.g., Ferrando, 2015). We have also tried to provide realistic choices in terms of sample sizes, test lengths, distributions of item/person parameters, and proportion of inconsistent respondents.

The conditions that were kept constant in all the simulations were the following: (a) the item scores were 5-point Likert scored as 1–5; (b) the intercepts μ_j were randomly and uniformly distributed between 1.5 and 4.5; and (c) the loadings λ_j were randomly and uniformly distributed between 0.3 and 0.8. As for the rationale of these choices, first, there seems to be agreement that five is the minimum number of categories from which linear FA can be considered to be a reasonable approximation (Ferrando, 2009; Rhemtulla et al., 2012). Second, condition (b) reflects a desirable condition in a general-purpose test: a wide range of difficulties evenly distributed. Finally, conditions (c) and (d) aim to reflect the results we generally find in FA applications in the personality domain.

Independent Variables

The study was based on a 2 × 3 × 3 × 4 × 7 design with the following independent variables: (a) sample size ($N = 500$ and $N = 1000$); (b) test length ($n = 20$, $n = 40$, $n = 60$); (c) percentage of inconsistent respondents (5, 15, 25%); (d) percentage of items in which responses were inconsistent (5, 10, 20, 30%), and (e) type of inconsistent responding. The seven types of simulated inconsistencies are described below.

1. Random responding (RAND). A very common type of misfit (Liu et al., 2016) expected in conditions of unmotivated responding and/or fatigue in the case of long tests. Responses

for the corresponding sub-set of items were generated using a random number generator.

2. Low person reliability (LPR) (e.g., Ferrando, 2015). Random responding can be considered as the extreme of a dimension of low person reliability characterized by a certain degree of insensitivity to the normative ordering of the items. This type of inconsistency was simulated here by generating the data according to Ferrando (2014) differential-discrimination model and setting the person parameter to a value of 0.20 for all of the item responses (a unit value is the expected value in the normative model).
3. Sabotaging (SAB). This is the tendency of the respondent to agree with the most extreme or "difficult" items and disagree with the "easier" items (see Ferrando, 2015). For the corresponding sub-set of items, responses at one extreme were changed to responses at the other extreme (e.g., 5–1 or 1–5).
4. Spuriously low unexpected responses (UE-L) and spuriously high unexpected responses (UE-H). There are expected to be inconsistencies of types UE-L and UE-H in some sub-sets of items mainly in the case of multidimensionality, faking (in the subset of socially desirable items), and acquiescence when balanced scales are used (see Ferrando, 2015). For the selected sub-set of items the expected central responses were moved one or two points down (spuriously low) or up (spuriously high).
5. Model-consistent extreme responding (EMC). Extreme responding was considered to be a general source of misfit that affects all items. In type EMC, the direction component of the response (agree-disagree) was model-based but the response was more extreme than expected from the model. This was simulated by moving responses to one or two points above the expected response in the model-expected direction.
6. Partially inconsistent extreme responding (EMIC). First, the simulation proceeded as for type EMC, but then the extreme responses were reversed for 20% of the items. So, for the majority of items the response behavior is model based, but for the remaining items it is "pure extreme responding" regardless of item content.

For RAND, SAB, UE-L, and UE-H, the conditions in (d) above apply. For LPR, EMC, and EMIC, inconsistency was simulated for all of the items, so the common percentage in (d) was 100%.

The general conditions described so far were considered for two scenarios. In the first, the structural (item) parameters μ_j,λ_j, and $\sigma^2_{\varepsilon j}$ were assumed to be known from previous calibrations, an "ideal" condition that is commonly used in IRT-based simulations. Although not implausible, this is not the usual situation in FA applications, and its main role here is to provide an upper benchmark for the effectiveness of the MB indices.

The second scenario is the most habitual in FA applications: neither the structural parameters nor the trait levels are known, and they are both estimated from the same sample by using the calibration-scoring procedure described above. Because (a) the item indices are now sample estimates, and (b) the sample contains a certain proportion of inconsistent respondents, the effectiveness of the indices must necessarily be lower than that of scenario 1. In all the conditions here, item calibration was based on Unweighted Least Squares (ULS) estimation for two reasons. First, ULS is quite robust and can be used with small-to-medium samples and relatively large models (Jöreskog, 2003), the most common situation in typical-response applications. Second, when the model to be fitted is not exactly correct but only an approximation (as discussed above), ULS tends to produce more accurate estimates than other theoretically superior procedures (e.g., Briggs and MacCallum, 2003).

Overall, the general design so far summarized had 684 different conditions. The number of replications per cell was 500.

Assessing the Effectiveness of the Indices

Effectiveness of a person-fit index can be defined as its ability to reliably detect disturbances of various types (e.g., Karabatsos, 2003). In this study, we are particularly interested in the seven types of disturbances described above. We used two approaches to assess effectiveness: the first studied the mean differences in the consistent and inconsistent groups, and the second, more graphical approach was based on Receiver Operating Curve (ROC) analysis.

In the first approach we used, Hedges's *g* effect size index as a simple summary measure. It was calculated for all the person-fit values and design cells. This index provides a general idea about the potential capability of the index for differentiating consistent and inconsistent respondents in an easily interpretable metric.

In the second approach, ROC curves were estimated and graphically displayed so that each graph showed (a) the curves corresponding to the five indices compared, (b) the diagonal line of no differentiation, and (c) the optimal operating point (defined below). As a summary of the ROC analysis we computed (a) the estimated area under the curve (AUC), and (b) the optimal operating point (OOP), which was estimated by using an un-informative prior. The first measure provides an overall summary of the index effectiveness. The second is of interest for suggesting plausible cutoff values. The ROC analysis was performed with the MATLAB Toolbox *perfcurve* routine (available at https://es.mathworks.com/help/stats/perfcurve.html).

RESULTS

General Results

In both scenarios, **Table 1** shows the overall results for the mean-comparison approach across all the conditions in the study. The table clearly reveals some general trends. As far as the MB indices are concerned, the means and standard deviations of *lico* and *licz* in the consistent groups (i.e., when the null hypothesis holds) are reasonably close to their expected values. It is also clear that, as expected, the effectiveness of the MB indices is substantially higher in scenario 1, and is especially high for r_{pm} and *lico*. In scenario 2, however, *lico* is the most effective MB index, and its effectiveness still seems to be good in this more realistic scenario.

We turn now to GB indices, which are only displayed once in **Table 1** because they do not depend on the model parameters. First, Bollen's *aii* is the least effective index, which was also expected because it was not designed specifically to detect inconsistent patterns but outliers in general. In contrast, r_{pg} shows a high amount of effectiveness and is the index

TABLE 1 | Mean-group comparisons: general results.

		GB		MB known parameters			MB sample calibration		
		Bollen's a_{ii}	**r_{pg}**	**$lico_i$**	**$licz_i$**	**r_{pm}**	**$lico_i$**	**$licz_i$**	**r_{pm}**
Inconsistent responses	$\bar{X}$	0.066	0.546	2.414	2.840	0.493	1.397	1.075	0.664
	S_X	0.042	0.260	1.720	3.610	0.329	0.614	1.187	0.185
Consistent responses	$\bar{X}$	0.051	0.756	1.001	0.120	0.778	0.940	−0.088	0.774
	S_X	0.032	0.075	0.268	1.013	0.070	0.268	0.890	0.071
Effect size (g)		0.447	1.77	1.99	1.62	1.99	1.35	1.25	1.17

that performs best when the structural parameters have to be estimated from the sample.

The ROC results for the mean-comparison results discussed so far are in **Figure 2**. **Figure 2A** shows the results for scenario 1 and **Figure 2B** for scenario 2. The results are in close agreement with those in **Table 1** (the correlation between effect size and the AUC is 0.94). Note that Bollen's index is not far from the diagonal line of no differentiation, and r_{pg} is the furthest from it. Note also that in **Figure 2A** *lico* and *licz* completely overlap, whereas in **Figure 2B** *lico* appears to be more effective than its standardized version, and overall is again the most effective MB index when items are sample-calibrated.

Specific Results

The results of the simulation are too numerous to be discussed here in detail. So, we shall provide only a summary of the most important of them. Full results are available from the authors.

We start with the non-significant results. There were no noticeable differences regarding sample size for any of the indices, possibly because a sample of $N = 500$ is large enough to provide stable results.

Figure 3 shows the effect-size estimates of effectiveness plotted against the seven different types of inconsistency. For clarity, Bollen's a_{ii} has been omitted, and only the sample-calibrated results are presented for *lico, licz*, and r_{pm}.

First, as expected, *lico* and *licz* have very similar profiles. However, as suggested by **Table 1**, **Figure 2**, the effectiveness of *lico* appears to be slightly but consistently higher than its standardized version. Second, the profiles of the correlational indices are similar one to another except for the fact that the simple r_{pg} considerably outperforms r_{pm} in EMIC and SAB. Finally, as also found in **Table 1**, **Figure 2**, r_{pg} is the most effective index overall (when item parameters are not known). However, it is not consistently superior, and *lico* appears to be more effective in LPR and, above all, EMC, as was predicted above. Taking into account all the results so far, a reasonable choice for practical applications would be a combination of *lico* and r_{pg}. And these are the only indices that we shall consider from now on.

For the two indices selected, **Figure 4** shows the effect-size estimates of effectiveness plotted against test length. In both cases, effectiveness increases with the number of items. It is generally higher for r_{pg} but there tends to be fewer differences between them as the test becomes longer, and, furthermore, these differences are rather small in AUC units. In the 60-item condition, the results in **Figure 4** correspond to an AUC of 0.82 for both indices, which means a respectable amount of effectiveness. At the other extreme, for 20 items the AUCs would be of 0.74 for *lico* and 0.76 for r_{pg}, which are relatively low. Overall then, the results are similar for both indices, and agree with what has been reported in the person-fit literature: practical indices are generally ineffective in short tests of fewer than 20 items, and effectiveness increases mostly as a function of test length (e.g., Ferrando, 2015).

Figure 5 displays effect size against the percentage of inconsistent respondents, and results are again in accordance with the person-fit literature: effectiveness decreases as the proportion of inconsistent individuals increases. Note also that the decrease is more pronounced for *lico*, and that this index would be expected to be more effective than r_{pg} when the proportion of inconsistent respondents is low: at the 5% level, the AUC of *lico* is 0.90 against a 0.82 value for r_{pg}.

Finally, **Figure 6** displays effect size against the percentage of items in which inconsistent responses were given. It is in this condition that the two indices differ most. The effectiveness of r_{pg} clearly increases with the proportion of inconsistent items while the effectiveness of *lico* tends to decrease. Furthermore, at the 30% level the difference in terms of AUC is considerable: 0.76 for *lico* against 0.98 for r_{pg}. The most plausible explanation for this divergent behavior is the MB vs. GB nature of both indices: as the proportion of inconsistent items increases, the item parameter estimates at the calibration stage become more and more degraded, and this, in turn, decreases the effectiveness of the person-fit index via the mechanism explained below. In ‘ort of this explanation, we note that, when the item parameters are known (scenario 1), the trends of *lico* and r_{pg} in this condition are the same.

Cutoff Values

For the sample-based *lico*, **Table 2** shows the empirical standard deviations, the approximate expected standard deviations given by *sqrt(2/(n-1))*, and the cutoff values obtained from (a) the OOPs, and (b) Equation (11).

The results in **Table 2** are interesting. First, the empirical standard deviations decrease with test length, just as it should be, and agree rather well with their expected values. Second, the

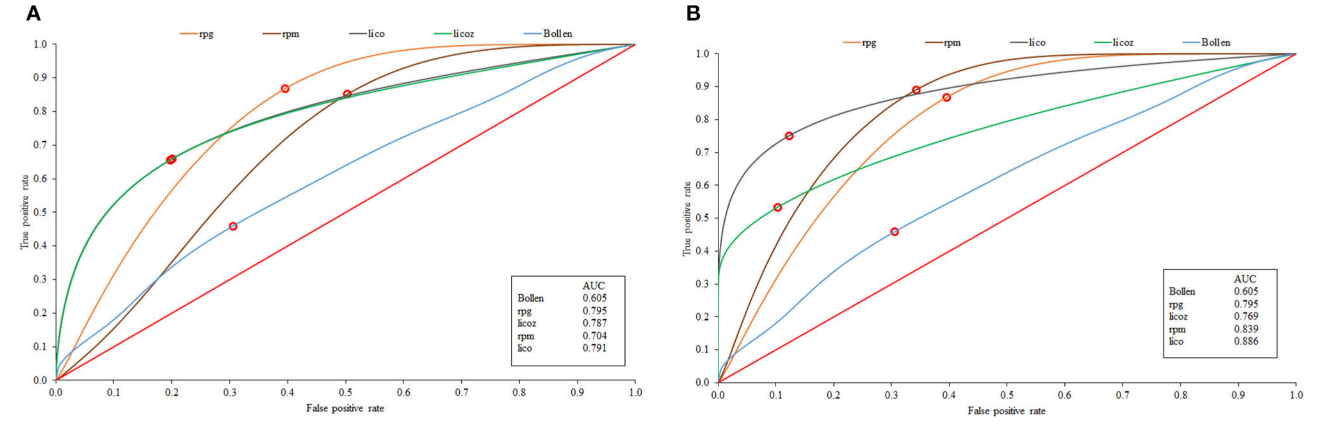


FIGURE 2 | ROC curves corresponding to the generals results for the known-parameter scenario (A), and sampled-calibrated items **(B)**.

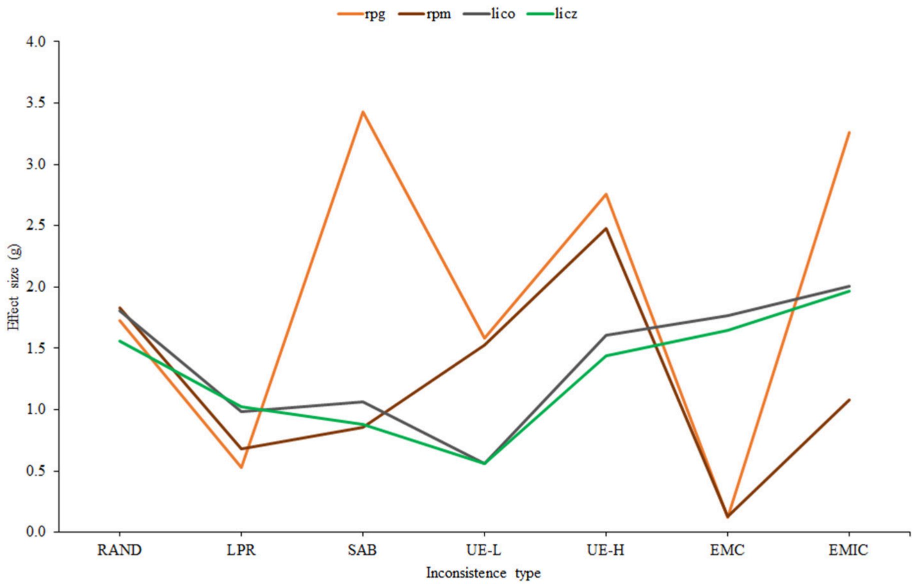


FIGURE 3 | Effect size estimates corresponding to the selected indices across different types of inconsistency.

simple cutoff values in Equation (11) proposed by Smith et al. (1998) are quite close to the OOPs when the item parameters are taken as known. For the sample-calibration case, however, cutoff values determined by 1+ *sqrt(2/(n-1)* (i.e., expected mean plus one expected standard deviation) will be closer to the corresponding OOPs. To sum up, it appears that simple cutoff values that only depend on test length can be proposed for practical applications based on *lico*. And further, the conventional 1.3–1.5 values proposed in Rasch modeling as a plausible general cutoff would possibly work reasonably well in practice.

ILLUSTRATIVE EXAMPLE

The short example provided in this section uses empirical data collected in personality research, and aims to (a) illustrate how the proposal made in the article can be used in practical applications, and (b) obtain further information regarding the behavior of the two chosen indices in real datasets when the conditions for effective person-fit assessment are far from ideal.

An 18-item Spanish version of Ray's balanced dogmatism scale (BDS, see Ferrando et al., 2016) was administered to a group of 346 undergraduate students. The items of this scale used a 6-point Likert format ranging from "completely disagree" (1) to completely agree (6).

First, the unidimensional FA model was fitted to the data using robust ULS estimation as implemented in version 10.4 of the FACTOR program (Lorenzo-Seva and Ferrando, 2013). The fit of the model was reasonably good (details can be obtained from the authors). Next the structural parameter estimates (μ,λ, and σ^2) were taken as fixed and known values, and (a) the ML trait

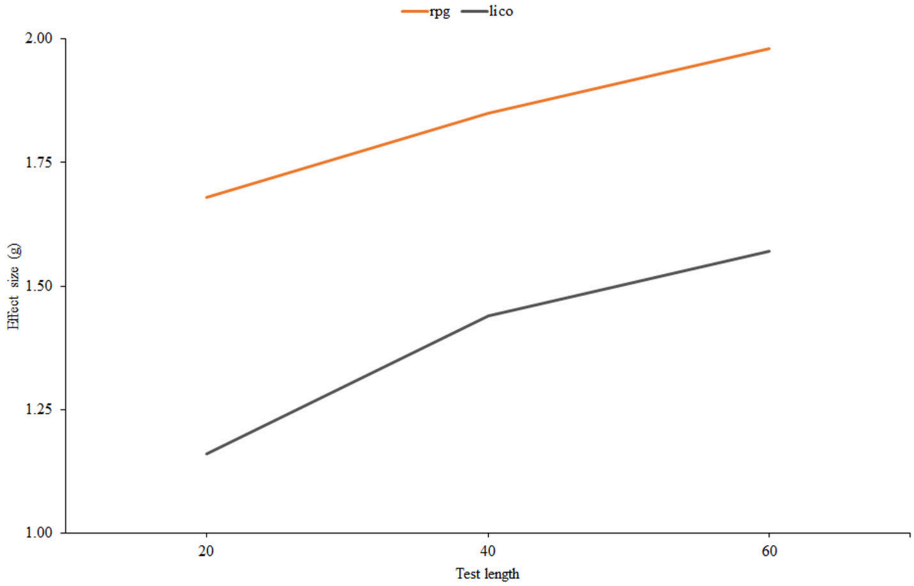


FIGURE 4 | Effect size estimates for r_{pg} and *lico* as related to test length.

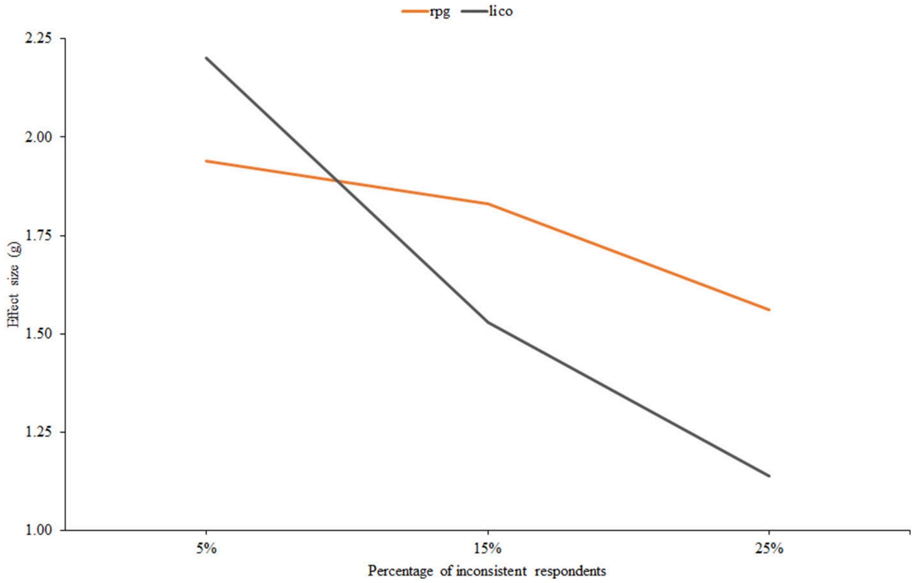


FIGURE 5 | Effect size estimates for r_{pg} and *lico* as related to the percentage of inconsistent respondents.

estimates and (b) the two indices proposed in this article were obtained using the new procedures implemented in FACTOR.

Inspection of the BDS item scores revealed that the items of the scale were "medium" to "easy," with means ranging from 3.08 to 5.73 (recall that the possible range of scores is 1–6). The lack of a wider range of item difficulties clearly diminishes the effectiveness of any person-fit measure (Ferrando, 2015), but is expected to have particular impact on the functioning of r_{pg} (see Equation 15). As the variability of the vector of item means decreases, the expected r_{pg} value approaches zero and the estimate becomes more unstable. This prediction was supported by the results: the mean value of r_{pg} in the sample was 0.53, lower than the usual values obtained in the simulation. The correlation between r_{pg} and *lico* was −0.41, which goes in the expected direction (see Equation 14) and indicates a moderate degree of agreement between both measures that would have been expected to be higher if the range of item difficulties had been wider. Finally, r_{pg} was obtained for all the respondents, which means that no single-category respondents appeared in the data. Overall, and in spite of the less than ideal

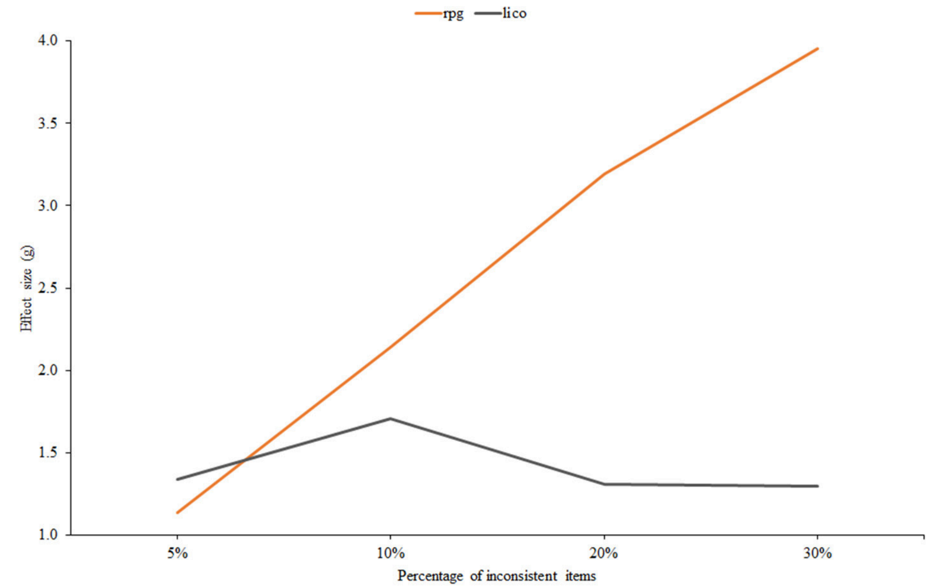


FIGURE 6 | Effect size estimates for r_{pg} and *lico* as related to the percentage of inconsistent items.

TABLE 2 | Standard deviation and cutoff values for *lico* as related to test length.

Num. items	Sd-Known parameters	Sd-sample parameters	Expected Sd	OOP Known parameters	OOP-sample calibration	$1+\frac{2}{\sqrt{n}}$
20	0.346	0.336	0.32	1.381	1.141	1.447
40	0.238	0.224	0.23	1.292	1.129	1.316
60	0.197	0.210	0.18	1.279	1.121	1.258

conditions r_{pg} is still expected to be useful here as an auxiliary index.

Lico seemed to work rather well even in these relatively unfavorable conditions (short test with a reduced range of item difficulties). The mean value of *lico* was 0.99 (virtually its expected value) and the corresponding standard deviation was 0.49, which is somewhat above the expected value of 0.34 (approximate) for *sqrt(2/(n-1))*. This result is only to be expected if the sample contains a certain proportion of inconsistent respondents. The distribution of the *lico* values can be seen in **Figure 7**.

The right tail of the distribution in **Figure 1** presumably contains those subjects who responded inconsistently with the FA model and with whom we wish to identify. We used Smith's critical value in Equation (11) (i.e., 1.47 with 11 items) and flagged 55 respondents (16% of the sample) as potentially inconsistent. Inspection of the corresponding patterns using the procedures proposed by Ferrando and Lorenzo-Seva (2016) suggested that the main sources of inconsistency were: (a) model-based extreme responding (characterized by a high value of *lico* and an above-average value of r_{pg}), (b) unexpected responses to certain sub-sets of items (in this case r_{pg} was generally low), and, to a lesser extent, (c) random responding/low person reliability (characterized by a near zero value of r_{pg}). Two possible cases of sabotaging or malingering (characterized by a high value of *lico* and a strong negative r_{pg}value) were also identified.

SUMMARY, PROPOSAL AND IMPLEMENTATION

The results described so far suggest that the combined use of *lico* and r_{pg} would be an effective first-step approach for flagging potentially inconsistent respondents in applications based on the standard FA model. The indices selected show a different profile of effectiveness across different types of inconsistency (**Figure 3**), and they also behave differently in terms of the proportion of items which are answered inconsistently (**Figure 6**). As for similarities, both essentially depend on the general conditions that affect person-fit indices (Ferrando, 2015): their effectiveness mostly depends on test length and decreases as the proportion of inconsistent respondents increases. Furthermore, the results of the empirical example show that a reduced range of item difficulties diminishes the effectiveness of the indices, especially that of r_{pg}. They also show, however, that even in the case of a relative short test with a reduced range of difficulty, the proposed indices work reasonably well. Overall, we believe that in a test with a minimal length of about 25 items and in which the proportion of inconsistent respondents is relatively small (say <10%), the approach proposed here would be expected to be highly effective in practice.

As discussed below, we do not feel that the present results allow strict cutoff values to be proposed for the selected indices.

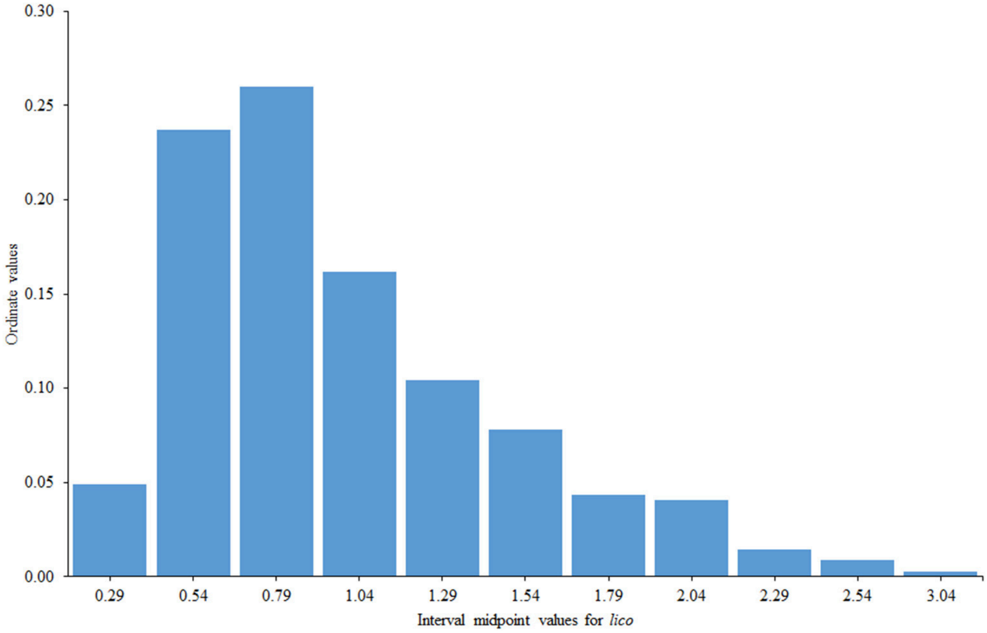


FIGURE 7 | Distribution of *lico* values in the illustrative example.

The expected values of r_{pg} depend on too many factors, so it seems better to use it as an auxiliary index, as proposed. As for *lico*, the cutoff values in Equation (11) seem to work reasonably well, but the results of the illustrative study suggest that they might even be too sensitive (16% of inconsistent respondents in a sample of volunteers seems to be a bit too high).

Because the results of the study are encouraging and a workable proposal can be made, the indices chosen and the reference values discussed above have been implemented in version 10.4 of the program FACTOR (Lorenzo-Seva and Ferrando, 2013), a free, comprehensive program for fitting the FA model. Furthermore, Matlab functions and illustrative data are offered as Supplementary Material.

DISCUSSION

Simple and effective practical indices based on the linear FA model can be used and easily implemented (as they have been) in a standard FA program. These are the main conclusions of the present study. At the same time, however, the study does have some limitations, and the results point out issues that can be improved or that deserve further research.

We shall start with a caveat. Practical person-fit indices are non-specific screening devices for tracing potentially inconsistent respondents. Ideally, however, once a pattern has been flagged as potentially inconsistent, further information should be obtained regarding (among other things) (a) the type of inconsistency, and (b) the impact that the inconsistency has on the trait estimates (Emons et al., 2005). FA-based analytical and graphical procedures for obtaining this information already exist and are implemented in stand-alone programs (Ferrando and Lorenzo-Seva, 2016). The problem may be how to link the first-step results obtained with a general FA program to this second-step type of analysis.

An alternative approach to using a first-step practical index followed by a second-step *post-hoc* analysis is to include the expected sources of misfit directly in the model (if this information is available). As far as we know, to date proposals of this type intended for the FA framework have been made for three sources of misfit: person unreliability (Ferrando, 2011), model-based extreme responding (Ferrando, 2014), and acquiescent responding (Ferrando et al., 2016). In this alternative approach, the use of the practical indices we propose has a secondary but important role that deserves further research: to detect the remaining inconsistent response patterns once the main expected sources of misfit have been explicitly taken into account in the model.

We turn now to more specific limitations and potential improvements. One clear limitation is that the study is only concerned with unidimensional FA solutions. In principle, our proposal is expected to work well not only with essentially unidimensional measures, but also with multidimensional instruments analyzed on a scale-by-scale basis, and instruments that behave according to a dominant factor solution (e.g., those that can be fitted with a bi-factor solution (see Reise, 2012). Even so, we acknowledge that many typical-response instruments are truly multidimensional questionnaires.

Since the personal correlation r_{pg} is a GB index, it can be obtained with no need to fit the FA model, and so it can be applied directly regardless of the number of factors. As for *lico*, its multidimensional extension is straightforward. So, the problem is not whether the indices generalize to the multidimensional case, but rather whether in this case they will be as effective as in the unidimensional setting. This point clearly requires further research.

The effectiveness of *lico* decreases when the trait estimates are poor (unreliable and/or indeterminate) and when the item parameters have to be estimated from the sample. These are important limitations. As for the first issue, we recommend checking the general quality of the trait estimates first by using marginal reliability measures and measures of factor indeterminacy such as Guttman's index (Guttman, 1955), before starting person-fit analysis. As for the second problem, Nering (1997) suggested one possible solution based on a two-stage calibration process in which (a) initial calibrations were run to identify potentially inconsistent patterns, (b) these patterns were removed from the data, and (c) items were recalibrated in the "cleaned" sample. It will be worth trying procedures of this type to see if levels of effectiveness can be obtained that are close to those achieved in the known-parameters scenario.

Finally, further research on cutoff values could be of interest. The simple cutoff criteria in Equation (11) considered here appear to work reasonably well as a starting point, but further study is required, and future substantive applications could also help to refine the proposal. On the other hand, person-based cutoff values obtained for each pattern using simulation (van Krimpen-Stoop and Meijer, 1999) could be a better alternative. Although they do require additional intensive computation, they are otherwise easily implemented.

In spite of the limitations discussed so far, we believe that what we propose here is a useful tool that allows the practitioner to routinely assess person fit in FA-based psychometric applications. As discussed above, this type of assessment is of considerable importance, so we hope that our proposal will be widely used in the near future.

AUTHOR CONTRIBUTIONS

PF initiated the paper, advised on simulation conditions and the choice of indices tested, and coordinated team meetings. AV proofread and provided recommendations. UL conducted the simulation studies and summarized the outcomes.

FUNDING

The research was supported by a grant from the Catalan Ministry of Universities, Research and the Information Society (2014 SGR 73), and by a grant from the Spanish Ministry of Education and Science (PSI2014-52884-P).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpsyg.2016.01973/full#supplementary-material

The MATLAB functions to compute the selected indices (*lico* and r_{pg}) as well as the data corresponding to the illustrative example are made available to the interested researchers.

REFERENCES

Armstrong, R. D., Stoumbos, Z. G., Kung, M. T., and Shi, M. (2007). On the performance of l_z statistic in person fit measurement. *Pract. Assess. Res. Eval.* 12. doi: 10.1177/01466216980221004

Bollen, K. A. (1987). Outliers and improper solutions a confirmatory factor analysis example. *Sociol. Methods Res.* 15, 375–384. doi: 10.1177/0049124187015004002

Bollen, K. A., and Arminger, G. (1991). "Observational residuals in factor analysis and structural equation models," in *Sociological Methodology 1991,* ed P. V. Marsden (New York, NY: Basil Blackwell), 235–262.

Briggs, N. E., and MacCallum, R. C. (2003). Recovery of weak common factors by maximum likelihood and ordinary least squares estimation. *Multivariate Behav. Res.* 38, 25–26. doi: 10.1207/S15327906MBR3801_2

Clark, J. M. (2010). *Aberrant Response Patterns as a Multidimensional Phenomenon: using Factor-Analytic Model Comparison to Detect cheating.* Doctoral dissertation, University of Kansas.

Conijn, J. M., Emons, W. H. M., De Jong, K., and Sitjsma, K. (2015). Detecting and explaining aberrant responding to the outcome questionnaire-45. *Assessment* 22, 513–524. doi: 10.1177/1073191114560882

Conijn, J. M., Emons, W. H. M., and Sijtsma, K. (2014). Statistic lz-based person-fit methods for noncognitive multiscale measures. *Appl. Psychol. Meas.* 38, 122–136. doi: 10.1177/0146621613497568

Cronbach, L. J., and Gleser, G. C. (1953). Assessing similarity between profiles. *Psychol. Bull.* 50, 456. doi: 10.1037/h0057173

Culpepper, S. A. (2013). The reliability and precision of total scores and IRT estimates as a function of polytomous IRT parameters and latent trait distribution. *Appl. Psychol. Meas.* 37, 201–225. doi: 10.1177/0146621612470210

Donlon, T. F., and Fischer, F. E. (1968). An index of an individual's agreement with group-determined item difficulties. *Educ. Psychol. Meas.* 28, 105–113. doi: 10.1177/001316446802800110

Emons, W. H. M., Sijtsma, K., and Meijer, R. R. (2005). Global, local and graphical person-fit analysis using person-response functions. *Psychol. Methods* 10, 101–119. doi: 10.1037/1082-989X.10.1.101

Ferrando, P. J. (2007). Factor-analytic procedures for assessing response pattern scalability. *Multivariate Behav. Res.* 42, 481–508. doi: 10.1080/00273170701382583

Ferrando, P. J. (2009). Difficulty, discrimination and information indices in the linear factor-analytic model for continuous responses. *Appl. Psychol. Meas.* 33, 9–24. doi: 10.1177/0146621608314608

Ferrando, P. J. (2011). A linear variable-θ model for measuring individual differences in response precision. *Appl. Psychol. Meas.* 35, 200–216. doi: 10.1177/0146621610391649

Ferrando, P. J. (2014). A factor-analytic model for assessing individual differences in response scale usage. *Multivariate Behav. Res.* 49, 390–405. doi: 10.1080/00273171.2014.911074

Ferrando, P. J. (2015). "Assessing person fit in typical-response measures," in *Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment,* eds S. P. Reise and D. A. Revicki (New York, NY: Routledge/Taylor & Francis Group), 128–155.

Ferrando, P. J., and Lorenzo-Seva, U. (2013). *Unrestricted Item Factor Analysis and Some Relations with Item Response Theory*. Technical Report, Department of Psychology, Universitat Rovira i Virgili, Tarragona. Available online at: http://psico.fcep.urv.es/utilitats/factor

Ferrando, P. J., and Lorenzo-Seva, U. (2016). A comprehensive regression-based approach for identifying sources of person misfit in typical-response measures. *Educ. Psychol. Meas.* 76, 470–486. doi: 10.1177/0013164415594659

Ferrando, P. J., Morales-Vives, F., and Lorenzo-Seva, U. (2016). Assessing and controlling acquiescent responding when acquiescence and content are related: a comprehensive factor-analytic approach. *Struct. Equation Model.* 23, 713–725, doi: 10.1080/10705511.2016.1185723

Fowler, H. M. (1954). An application of the Ferguson method of computing item conformity and person conformity. *J. Exp. Educ.* 22, 237–245. doi: 10.1080/00220973.1954.11010480

Guttman, L. (1955). The determinacy of factor score matrices with implications for five other basic problems of common-factor theory 1. *Br. J. Stat. Psychol.* 8, 65–81. doi: 10.1111/j.2044-8317.1955.tb00321.x

Hofstee, W. K. B., Ten Berge, J. M. F., and Hendricks, A. A. J. (1998). How to score questionnaires. *Pers. Individ. Dif.* 25, 897–910. doi: 10.1016/S0191-8869(98)00086-5

Jensen, D. R., and Solomon, H. (1972). A Gaussian approximation to the distribution of a definite quadratic form. *J. Am. Stat. Assoc.* 67, 898–902. doi: 10.1080/01621459.1972.10481313

Jöreskog, K. G. (2003). *Factor Analysis by Minres*. Available online at: http://www.ssicentral.com/lisrel/techdocs/minres.pdf

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Appl. Meas. Educ.* 16, 277–298. doi: 10.1207/S15324818AME1604_2

Liu, T., Lan, T., and Xin, T. (2016). Detecting random responses in a personality scale using IRT-based person-FIT indices. *Eur. J. Psychol. Assess.* doi: 10.1027/1015-5759/a000369. [Epub ahead of print].

Lorenzo-Seva, U., and Ferrando, P. J. (2013). FACTOR 9.2: a comprehensive program for fitting exploratory and semiconfirmatory factor analysis and IRT models. *Appl. Psychol. Meas.* 37, 497–498. doi: 10.1177/0146621613487794

McDonald, R. P. (1982). Linear versus models in item response theory. *Appl. Psychol. Meas.* 6, 379–396. doi: 10.1177/014662168200600402

Meijer, R. R., Niessen, A. S. M., and Tendeiro, J. N. (2015). A practical guide to check the consistency of item response patterns in clinical research through person-fit statistics examples and a computer program. *Assessment* 23, 52–62. doi: 10.1177/1073191115577800

Nering, M. L. (1997). Trait level estimation for nonfitting response vectors. *Appl. Psychol. Meas.* 21, 321–336. doi: 10.1177/01466216970214003

Reise, S. P. (1995). Scoring method and the detection of person misfit in a personality assessment context. *Appl. Psychol. Meas.* 19, 213–229. doi: 10.1177/014662169501900301

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behav. Res.* 47, 667–696. doi: 10.1080/00273171.2012.715555

Reise, S. P., and Widaman, K. F. (1999). Assessing the fit of measurement models at the individual level: a comparison of item response theory and covariance structure approaches. *Psychol. Methods* 4, 3–21. doi: 10.1037/1082-989X.4.1.3

Rhemtulla, M., Brosseau-Liard, P. E., and Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robustcontinuous and categorical SEM estimation methods under suboptimal conditions. *Psychol. Methods* 17, 354. doi: 10.1037/a0029315

Rudner, L. M. (1983). Individual assessment accuracy. *J. Educ. Meas.* 20, 207–219. doi: 10.1111/j.1745-3984.1983.tb00200.x

Rupp, A. A. (2013). A systematic review of the methodology for person fit research in item response theory: Lessons about generalizability of inferences from the design of simulation studies. *Psychol. Test Assess. Model.* 55, 3–38.

Smith, R. M. (1986). Person fit in the Rasch model. *Educ. Psychol. Meas.* 46, 359–372. doi: 10.1177/001316448604600210

Smith, R. M., Schumacher, R. E., and Bush, M. J. (1998).Using item mean squares to evaluate fit to the Rasch model. *J. Outcome Meas.* 2, 6–78.

Tendeiro, J. N., and Meijer, R. R. (2014). Detection of invalid test scores: the usefulness of simple nonparametric statistics. *J. Educ. Meas.* 51, 239–259. doi: 10.1111/jedm.12046

International Test Commission (2014). *ITC Guidelines for Quality Control in Scoring, Test Analysis, and Reporting of Test Scores*. Available online at: http://intestcom.org (Accessed February 25, 2014)

van Krimpen-Stoop, E. M. L. A., and Meijer, R. R. (1999). The null distribution of person-fit statistics for conventional and adaptive tests. *Appl. Psychol. Meas.* 23, 327–345. doi: 10.1177/01466219922031446

Wilson, E. B., and Hilferty, M. M. (1931). The distribution of chi-square. *Proc. Natl. Acad. Sci. U.S.A.* 17, 684–688. doi: 10.1073/pnas.17.12.684

Wright, B. D., and Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Meas. Trans.* 8, 370.

Wright, B. D., and Masters, G. N. (1982). *Rating Scale Analysis. Rasch Measurement*. Chicago, IL: MESA press.

Wright, B. D., Mead, R. J., and Bell, S. R. (1979). *BICAL: Calibrating Items with the Rasch Model*. Statistical Laboratory, Department of Education, University of Chicago.

Yuan, K. H., Fung, W. K., and Reise, S. P. (2004). Three Mahalanobis distances and their role in assessing unidimensionality. *Br. J. Math. Stat. Psychol.* 57, 151–165. doi: 10.1348/000711004849231

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 11 May 2017
doi: 10.3389/fpsyg.2017.00747

Comparing Indirect Effects in Different Groups in Single-Group and Multi-Group Structural Equation Models

Ehri Ryu[1]* *and Jeewon Cheong*[2]

[1] *Psychology, Boston College, Chestnut Hill, MA, USA,* [2] *Health Education and Behavior, University of Florida, Gainesville, FL, USA*

In this article, we evaluated the performance of statistical methods in single-group and multi-group analysis approaches for testing group difference in indirect effects and for testing simple indirect effects in each group. We also investigated whether the performance of the methods in the single-group approach was affected when the assumption of equal variance was not satisfied. The assumption was critical for the performance of the two methods in the single-group analysis: the method using a product term for testing the group difference in a single path coefficient, and the Wald test for testing the group difference in the indirect effect. Bootstrap confidence intervals in the single-group approach and all methods in the multi-group approach were not affected by the violation of the assumption. We compared the performance of the methods and provided recommendations.

Keywords: moderated mediation, moderated indirect effect, group difference in mediation, multi-group analysis, simple indirect effect

OPEN ACCESS

Edited by:
Prathiba Natesan,
University of North Texas, USA

Reviewed by:
Jiun-Yu Wu,
National Chiao Tung University, Taiwan
Victoria Savalei,
University of British Columbia, Canada

Correspondence:
Ehri Ryu
ehri.ryu.1@bc.edu

Specialty section:
This article was submitted to Quantitative Psychology and Measurement,
a section of the journal Frontiers in Psychology

Received: *23 December 2016*
Accepted: *24 April 2017*
Published: *11 May 2017*

Citation:
Ryu E and Cheong J (2017) Comparing Indirect Effects in Different Groups in Single-Group and Multi-Group Structural Equation Models. Front. Psychol. 8:747.
doi: 10.3389/fpsyg.2017.00747

INTRODUCTION

In mediation analysis, it is a standard practice to conduct a formal statistical test on mediation effects in addition to testing each of the individual parameters that constitutes the mediation effect. Over the past few decades, statistical methods have been developed to achieve valid statistical inferences about mediation effects. The sampling distribution of a mediation effect is complicated because the mediation effect is quantified by a product of at least two parameters. For this reason, numerous studies have proposed and recommended methods that do not rely on distributional assumption (e.g., bootstrapping) for testing mediation effects (e.g., Bollen and Stine, 1990; Shrout and Bolger, 2002; MacKinnon et al., 2004; Preacher and Hayes, 2004).

It is often a question of interest whether a mediation effect is the same across different groups of individuals or under different conditions, in other words, whether a mediation effect is moderated by another variable (called a moderator) that indicates the group membership or different conditions. For example, Levant et al. (2015) found that the mediation effect of endorsement of masculinity ideology on sleep disturbance symptoms via energy drink use was significantly different between white and racial minority groups. Schnitzspahn et al. (2014) found that time monitoring mediated the effect of mood on prospective memory in young adults, but not in old adults. Gelfand et al. (2013) showed that the effect of cultural difference (US vs. Taiwan) on the optimality of negotiation outcome is mediated by harmony norm when negotiating as a team but not when negotiating as solos. In these studies, the mediation effect was moderated by a categorical moderator (e.g., racial group, age group, experimental condition). With a categorical moderator, the moderated mediation effect concerns the difference in the indirect effect between groups. Treating a moderator categorical is appropriate when the moderator is truly categorical, but it is not appropriate to create groups based on arbitrary categorization of a continuous moderator (Maxwell and Delaney, 1993; MacCallum et al., 2002; Edwards and Lambert, 2007; Rucker et al., 2015).

Structural equation modeling (SEM) is a popular choice for many researchers to test a mediation model and to conduct a formal test on mediation effects. In SEM, the mediation effect can be specified as an indirect effect (Alwin and Hauser, 1975; Bollen, 1987) such as "the indirect effect of an independent variable (X) on a dependent variable (Y) via a mediator (M)" in which X affects M, which in turn affects Y. For incorporating a categorical moderator, there are two approaches in SEM: single-group and multi-group analysis. In the single-group analysis approach, the categorical moderator is represented by a variable, or a set of variables, in the model. On the other hand, the multi-group analysis approach uses the categorical moderator to separate the observations into groups at each level of the moderator, and the moderator does not appear in the model as a variable.

In this article, we present the single-group and multi-group analysis approaches to comparing indirect effects between groups, and introduce statistical methods in each approach for testing the group difference in the indirect effect and for testing the simple indirect effect in each group. Then we present a simulation study to compare the performance of the methods. In particular, we examine how robust the methods in single-group analysis approach are when the assumption of homogeneity of variance is not satisfied (the assumption is described in a later section).

GROUP DIFFERENCE IN INDIRECT EFFECT AND SIMPLE INDIRECT EFFECT IN EACH GROUP

We use the following example throughout this article. Suppose that we hypothesize a mediation model in which the effect of an independent variable X on a dependent variable Y is mediated by a mediator M (**Figure 1**).

We also hypothesize that the X to M relationship is not the same in two groups of individuals (e.g., men and women). This model can be considered as a special case of the first *stage moderation model* in Edwards and Lambert (2007) and the *Model 2* in Preacher et al. (2007), in which the moderator is a categorical variable with two levels. When comparing the indirect effect between two groups, estimating and making statistical inferences about the following two effects are of interest. First, what is the estimated difference in the indirect effect between the groups? Second, what is the estimated indirect effect in each group (i.e., simple indirect effect)?

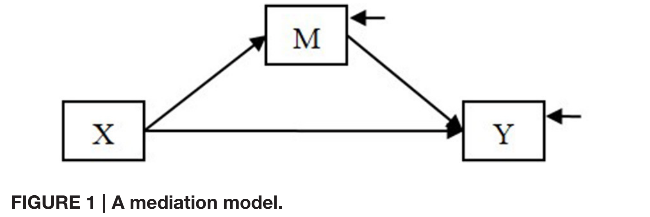


FIGURE 1 | A mediation model.

In the single-group analysis, a (set of) categorical variable indicating the group membership is used as a covariate in the model and an interaction term of X with the group membership (Group) is included to test the difference in the X to M relationship between groups (See **Figure 2A**).

The interpretation of the parameters depends on how the group membership is coded. For example, when the group membership (Group) is dummy coded as 1 = Group 1 and 0 = Group 2, a_1 = simple effect of X on M in Group 2; a_2 = group difference in conditional mean of M for those whose level of X is at zero (i.e., conditional mean of M in Group 1—conditional mean of M in Group 2); a_3 = difference in simple effect of X on M between groups (i.e., simple effect of X on M in Group 1—simple effect of X on M in Group 2). If $a_3 \neq 0$, it means that the relationship between X and M is not the same between groups.

When the relationship between X and M differs between groups, the indirect effect of X on Y via M is conditional on the group membership, because the indirect effect consists of X to M relationship and M to Y relationship. In the model shown in **Figure 2A**, an estimate of the indirect effect of X on Y via M is obtained by $\left[\hat{a}_1 + \hat{a}_3 \left(Group\right)\right] \hat{b}$ (Preacher et al., 2007). So the simple indirect effect (i.e., the conditional indirect effect)

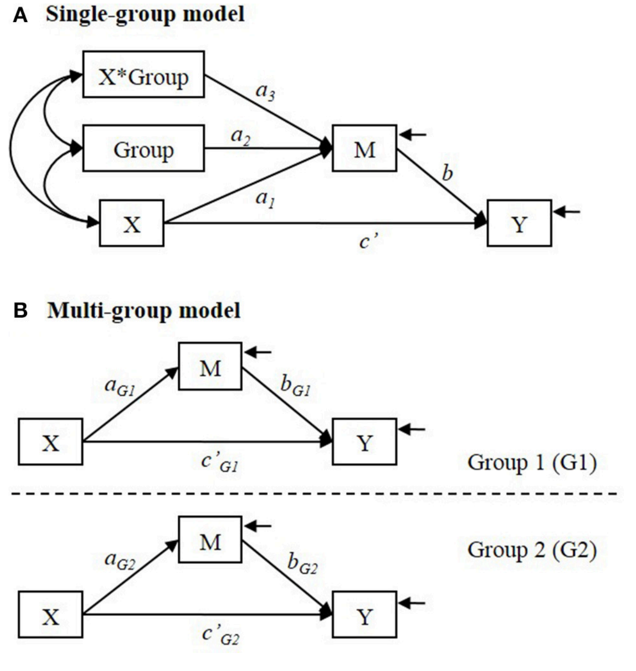


FIGURE 2 | (A) Single-group and **(B)** multi-group analysis models for testing group difference in the indirect effect. In **(A)** single-group model, Group is a categorical variable that indicates distinctive group membership.

TABLE 1 | Methods for testing group difference in a path, group difference in the indirect effect, and simple indirect effect in each group.

	Abbreviation	Description
SINGLE-GROUP ANALYSIS		
Group difference in *a* path	z^{S}_{a3}	$z = \hat{a}_3/se_{a3}$
Group difference in the indirect effect	W^{S}_{diff}	Wald test for $a_3 b = 0$
Simple indirect effect in each group	PC^{S}_{ind}	Percentile bootstrap CI for the simple indirect effect in each group
	BC^{S}_{ind}	Bias-corrected bootstrap CI for the simple indirect effect in each group
MULTI-GROUP ANALYSIS		
Group difference in *a* path	LR^{M}_{a}	Likelihood ratio test for $a_{G1} = a_{G2}$
Group difference in the indirect effect	LR^{M}_{diff}	Likelihood ratio test for $a_{G1}b_{G1} = a_{G2}b_{G2}$
	W^{M}_{diff}	Wald test for $a_{G1}b_{G1} = a_{G2}b_{G2}$
	PC^{M}_{diff}	Percentile bootstrap CI for the group difference in the indirect effect
	BC^{M}_{diff}	Bias-corrected bootstrap CI for the group difference in the indirect effect
	MC^{M}_{diff}	Monte Carlo CI for the group difference in the indirect effect
Simple indirect effect in each group	PC^{M}_{ind}	Percentile bootstrap CI for the simple indirect effect in each group
	BC^{M}_{ind}	Bias-corrected bootstrap CI for the simple indirect effect in each group
	MC^{M}_{ind}	Monte Carlo confidence interval for the simple indirect effect in each group

The superscripts "S" and "M" indicate the single-group and multi-group approaches, respectively. The subscript "ind" indicates the simple indirect effect in each group; the subscript "diff" indicates the group difference in the indirect effect. CI, confidence interval. We used 95% confidence for all interval estimates.

estimate is $\left[\hat{a}_1 + \hat{a}_3\,(1)\right]\hat{b} = \left(\hat{a}_1 + \hat{a}_3\right)\hat{b}$ in Group 1 (coded 1), and $\left[\hat{a}_1 + \hat{a}_3\,(0)\right]\hat{b} = \hat{a}_1\hat{b}$ in Group 2 (coded 0). The estimated group difference in the indirect effect is $\left[\left(\hat{a}_1 + \hat{a}_3\right)\hat{b}\right] - \hat{a}_1\hat{b} = \hat{a}_3\hat{b}$ (Hayes, 2015).

In multi-group analysis, group membership is not used as a predictor variable in the model. Instead, a set of hypothesized models (e.g., a set of two models if there are two distinctive groups) are specified and estimated simultaneously (See **Figure 2B**). The group difference in the simple effect of X on M (that is estimated by a_3 in the single-group analysis) is estimated by $\left(\hat{a}_{G1} - \hat{a}_{G2}\right)$. The simple indirect effect is estimated by $\hat{a}_{G1}\hat{b}_{G1}$ and $\hat{a}_{G2}\hat{b}_{G2}$ in Group 1 and in Group 2, respectively. The estimated difference in the indirect effect is $\left(\hat{a}_{G1}\hat{b}_{G1} - \hat{a}_{G2}\hat{b}_{G2}\right)$.

STATISTICAL INFERENCES

There are numerous methods for making statistical inferences about the simple indirect effects and inferences about the group difference in the indirect effect. The methods can be categorized into the following branches: (1) normal-theory standard error, (2) bootstrapping methods, (3) Monte Carlo method, (4) likelihood ratio (LR) test, (5) Wald test[1]. **Table 1** summarizes the methods and shows the abbreviation to refer to each method. In the abbreviation, the superscripts "S" and "M" indicate the single-group and multi-group approaches, respectively. The subscripts indicate which effect is tested by the method, e.g., "diff" means the group difference in the indirect effect, "ind" means the simple indirect effect in in each group.

Normal-Theory Standard Error

The normal-theory standard error method is based on the assumption that the sampling distribution of the estimate follows a normal distribution. In testing an indirect effect, it is well-known that the normality assumption is not appropriate to represent the sampling distribution of the indirect effect, and the normal-theory based method do not perform well in testing the indirect effect (e.g., MacKinnon et al., 2002; Shrout and Bolger, 2002; MacKinnon et al., 2004; Preacher and Selig, 2012). In moderated mediation models, Preacher et al. (2007) has advocated the bootstrapping methods over the normal standard error methods for testing the simple indirect effect.

Bootstrapping Methods

The bootstrapping methods can provide interval estimates without relying on a distribution assumption. For this reason, the bootstrapping methods have been recommended for testing indirect effects in previous studies (e.g., MacKinnon et al., 2004; Preacher and Hayes, 2004). The bootstrapping methods can be applied for obtaining interval estimates for any effect of interest, e.g., simple indirect effect in Group 1, simple indirect effect in Group 2, group difference in the indirect effect. In bootstrapping methods, a large number of bootstrap samples (e.g., 1,000 bootstrap samples), whose sizes are the same as the original sample size, are drawn from the original sample with replacement. An estimate is obtained in each bootstrap sample. An empirical sampling distribution is constructed using the set of 1,000 bootstrap estimates. From the bootstrap

[1] In mediation analysis, the poor performance of the method based on the normality assumption is well-known. We included the normal theory standard error method in the simulation study. As expected, and consistent with the previous findings in the literature, the normal standard error method did not perform well. We introduce the method here for the purpose of reviewing previous literature but do not consider the normal-theory standard error method hereafter to avoid redundancy. The normal-theory standard error does not appear in **Table 1**. We do not present simulation results regarding this method.

sampling distribution, percentile bootstrap confidence intervals ([100 * (1−α)]%) can be computed by the (α/2) and (1−α/2) percentiles. Bias-corrected bootstrap confidence intervals can be computed with the percentiles adjusted based on the proportion of bootstrap estimates lower than the original sample estimate (see MacKinnon et al., 2004).

In the single-group analysis, the estimate of the simple indirect effect in each group is computed by $\left(\hat{a}_1^* + \hat{a}_3^*\right)\hat{b}^*$ in Group 1 (coded 1), and $\hat{a}_1^*\hat{b}^*$ in Group 2 (coded 0) in each bootstrap sample. The superscript $*$ denotes that the estimates are obtained in bootstrap samples. In each group, the percentile (PC_{ind}^S in **Table 1**) and the bias-corrected (BC_{ind}^S) bootstrap confidence intervals for the simple indirect effect are computed from the bootstrap sampling distribution [i.e., the distribution of $\left(\hat{a}_1^* + \hat{a}_3^*\right)\hat{b}^*$ for Group 1; and the distribution of $\hat{a}_1^*\hat{b}^*$ for Group 2] as described above.

In the multi-group analysis, the estimate of the simple indirect effect is computed by $\hat{a}_{G1}^*\hat{b}_{G1}^*$ in Group 1 and $\hat{a}_{G2}^*\hat{b}_{G2}^*$ in Group 2. The percentile (PC_{ind}^M) and the bias-corrected (BC_{ind}^M) bootstrap confidence intervals for the simple indirect effect are obtained from the distribution of $\hat{a}_{G1}^*\hat{b}_{G1}^*$ and the distribution of $\hat{a}_{G2}^*\hat{b}_{G2}^*$, in Group 1 and Group 2, respectively. The percentile (PC_{diff}^M) and the bias-corrected (BC_{diff}^M) bootstrap confidence intervals for the group difference in the indirect effect are obtained from the bootstrap sampling distribution of $\left(\hat{a}_{G1}^*\hat{b}_{G1}^* - \hat{a}_{G2}^*\hat{b}_{G2}^*\right)$.

Monte Carlo Method

The Monte Carlo method provides a statistical test or an interval estimate of an effect by generating parameter values with a distributional assumption (e.g., multivariate normal). For testing the group difference in the indirect effect in the multi-group analysis model, the parameter estimates and standard errors are used to specify a joint sampling distribution of the parameter estimates from which the parameter values are generated for a large number of replications, e.g., 1,000 (Preacher and Selig, 2012; Ryu, 2015), such that the joint distribution of the four parameters a_{G1}, b_{G1}, a_{G2}, and b_{G2} is a multivariate normal distribution shown below.

$$\begin{bmatrix} a_{G1} \\ b_{G1} \\ a_{G2} \\ b_{G2} \end{bmatrix} \sim MVN\left(\begin{bmatrix} \hat{a}_{G1} \\ \hat{b}_{G1} \\ \hat{a}_{G2} \\ \hat{b}_{G2} \end{bmatrix}, \begin{bmatrix} \hat{\sigma}_{a_{G1}}^2 & & & \\ 0 & \hat{\sigma}_{b_{G1}}^2 & & \\ 0 & 0 & \hat{\sigma}_{a_{G2}}^2 & \\ 0 & 0 & 0 & \hat{\sigma}_{b_{G2}}^2 \end{bmatrix}\right) \quad (1)$$

where $\hat{a}_{G1}$, $\hat{b}_{G1}$, $\hat{a}_{G2}$, and $\hat{b}_{G2}$ are the estimates in the original sample, and $\hat{\sigma}_{a_{G1}}$, $\hat{\sigma}_{b_{G1}}$, $\hat{\sigma}_{a_{G2}}$, and $\hat{\sigma}_{b_{G2}}$ are the estimated standard errors in the original sample. The parameters in Group 1 (a_{G1}, b_{G1}) are independent of the parameters in Group 2 (a_{G2}, b_{G2}) because Group 1 and Group 2 are independent as long as the assumption of independent observations is valid. In mediation model, the covariance between a and b paths are often replaced with zero (Preacher and Selig, 2012). So the covariance between a and b paths is zero in each group ($\hat{\sigma}_{b_{G1},a_{G1}} = 0$; $\hat{\sigma}_{b_{G2},a_{G2}} = 0$). For a large number of replications, parameter values $\hat{a}_{G1}^+$, $\hat{b}_{G1}^+$, $\hat{a}_{G2}^+$, and $\hat{b}_{G2}^+$ are generated from the multivariate normal distribution shown in (1). The superscript $+$ denotes the parameter values generated by Monte Carlo method. In each replication, the simple indirect effect estimate is computed by $\hat{a}_{G1}^+\hat{b}_{G1}^+$ in Group 1 and by $\hat{a}_{G2}^+\hat{b}_{G2}^+$ in Group 2. The group difference in the indirect effect is computed by $\left(\hat{a}_{G1}^+\hat{b}_{G1}^+ - \hat{a}_{G2}^+\hat{b}_{G2}^+\right)$. The Monte Carlo confidence intervals ([100 * (1−α)]%) are obtained by the (α/2) and (1−α/2) percentiles in the set of generated values. For the simple indirect effect in Group 1, the Monte Carlo confidence intervals (MC_{ind}^M) are computed using the set of $\hat{a}_{G1}^+\hat{b}_{G1}^+$ values, and using the set of $\hat{a}_{G2}^+\hat{b}_{G2}^+$ values in each group, respectively. The Monte Carlo confidence interval for the group difference in the indirect effect (MC_{diff}^M) is obtained using the set of $\left(\hat{a}_{G1}^+\hat{b}_{G1}^+ - \hat{a}_{G2}^+\hat{b}_{G2}^+\right)$ values. The Monte Carlo method is less computer-intensive and less time-consuming than the bootstrapping method.

Likelihood Ratio Test

The likelihood ratio (LR) test and the Wald test can be used to test a (set of) constraint. The LR test (Bentler and Bonett, 1980; Bollen, 1989) is obtained by estimating two nested models with (M_1) and without (M_0) the constraints. The LR test results in a chi-square statistic with the degrees of freedom (df) equal to the difference in the number of freely estimated parameters in the two models.

$$\chi^2 = -2log\left[\frac{L(M_1)}{L(M_0)}\right] = \left\{-2log\left[L(M_1)\right]\right\} - \left\{-2log\left[L(M_0)\right]\right\} \quad (2)$$

where $L(M_k)$ = likelihood of model k. The LR test can be used to test the group difference in the "X → M" relationship in the multi-group analysis model, by comparing two models with and without the constraint $a_{G1} = a_{G2}$, with df = 1 (LR_a^M). Likewise, the LR test can be used to test the group difference in the indirect effect by comparing two models with and without the constraint $a_{G1}b_{G1} = a_{G2}b_{G2}$, with df = 1 ($LR_{diff}^M$).

Wald Test

The Wald test (Wald, 1943; Bollen, 1989) evaluates a constraint in a model in which the constraint is not imposed. For testing group difference in the indirect effect, the constraint $a_3b = 0$ is tested in the single-group analysis (W_{diff}^S). The Wald statistic (with df = 1) is obtained by

$$W = \hat{\theta}_1^2/avar\left(\hat{\theta}_1\right) \quad (3)$$

Where $\theta_1 = a_3b$ and $avar\left(\hat{\theta}_1\right)$ = estimated asymptotic variance of $\hat{\theta}_1$, i.e., estimated asymptotic variance of $\hat{a}_3\hat{b}$. Likewise, for testing group difference in the indirect effect in the multi-group model, the constraint $a_{G1}b_{G1} = a_{G2}b_{G2}$ is tested (W_{diff}^M). The Wald statistic (df = 1) is obtained by (3) with $\theta_1 = a_{G1}b_{G1} - a_{G2}b_{G2}$ in the multi-group model.

A previous simulation study (Ryu, 2015) compared the performance of different methods for testing group difference in the indirect effect in multi-group analysis. In the previous study, the LR test performed well in terms of Type I error rate and statistical power. The percentile bootstrap confidence intervals for the group difference in indirect effect showed coverage rates that are close to the nominal level. The bias-corrected bootstrap confidence intervals were more powerful than the percentile bootstrap confidence intervals but the bias-corrected bootstrap confidence intervals showed inflated Type I error rates.

SINGLE-GROUP AND MULTI-GROUP APPROACHES

The multi-group analysis model shown in **Figure 2B** is less restrictive the single-group analysis model shown in **Figure 2A**. In the single-group model shown in **Figure 2A**, b and c' paths are assumed to be equal between groups, whereas b and c' paths are allowed to differ between groups in the multi-group model, unless additional equality constraints are imposed. It is possible to specify a single-group model that allow b or c' paths to differ between groups. In order to allow these parameters to differ between groups in the single-group model, additional parameters need to be estimated or additional interaction terms need to be added. If the model shown in **Figure 2A** is modified by specifying the path coefficients "Group → Y" and "X*Group → Y" to be freely estimated, that will allow c' to differ between groups. In order to allow b to differ between groups, the model needs an additional variable "M*Group" and the path coefficients "Group → Y" and "M*Group → Y" need to be freely estimated. The multi-group model can be simplified by imposing equality constraints $\hat{b}_{G1} = \hat{b}_{G2}$ and / or $\hat{c}'_{G1} = \hat{c}'_{G2}$.

In the single-group model, the variance and covariance parameters are assumed to be equal as well, whereas in the multi-group model those parameters are not restricted to be the same between groups unless additional equality constraints are imposed. Specifically, in the single-group analysis model (as shown in **Figure 2A**) the residual variances of M and Y are assumed to be equal in both groups. The equal variance assumption in the single-group analysis is one of the standard assumptions in general linear models. The assumption is that the conditional variance of the dependent variable is homogeneous at all levels of the independent variables. For example, in regression analysis, the conditional variance of the dependent variable is assumed to be equal at all levels of the predictor variable. In between-subject analysis of variance or in t-test to compare two independent means, the within-group variance is assumed to be equal across all groups. It is well-known that the empirical Type I error rate can be different from the nominal level when the equal variance assumption is violated (e.g., Box, 1954; Glass et al., 1972; Dretzke et al., 1982; Aguinis and Pierce, 1998).

The purpose of this study is to introduce the single-group and multi-group approaches in SEM to comparing indirect effects between groups, and to empirically evaluate the performance of the statistical methods. Specifically, we aim to empirically evaluate how well the statistical methods (summarized in **Table 1**) perform for three questions in the moderated mediation model: (i) comparing the a path (X → M) between groups, (ii) comparing the indirect effect between groups, (iii) testing simple indirect effect in each group. The methods we considered are summarized in **Table 1**. We also evaluate how robust the methods in the single-group analysis are when the assumption of equal variances does not hold between groups. We expected that the performance of the methods in multi-group analysis would not be affected by the violation of the assumption of equal variances, because the multi-group analysis model does not rely on the assumption. In the single-group analysis, we expected that the performance of the z^{S}_{a3} and W^{S}_{diff} methods would be affected by the violation of the equal variance assumption, and that the confidence intervals produced by the bootstrapping methods (PC^{S}_{ind}, BC^{S}_{ind}) would not be affected by the violation of the assumption. The estimates are expected to be unbiased regardless of the equal variance assumption violated. The bootstrap sampling distribution is constructed using the estimates in bootstrap samples. Therefore, as long as the violation of the equal variance assumption does not affect the unbiasedness of the estimates, the performance of the bootstrap confidence intervals is not expected to be affected by the violation of the assumption.

SIMULATION

We used the mediation model shown in **Figure 2B** as the population model. There were two distinctive groups (denoted by G1 and G2). We considered a total of 63 conditions: 21 populations × 3 sample sizes.

As shown in **Table 2**, the 21 populations were created by combinations of three sets of parameter values for structural paths (Populations I, II, and III) and seven sets of parameter values for residual variances (Populations -0, -M1, -M2, -M3, -Y1, -Y2, -Y3). In Population I, there was no group difference in the indirect effect ($a_{G1}b_{G1} = 0.165$; $a_{G2}b_{G2} = 0.165$). In Population II, there was no indirect effect in G1; there was a small

TABLE 2 | Parameter values for structural paths *a* and *b*, and for residual variances of M and Y in population.

Population	Parameter values
PARAMETER VALUES FOR STRUCTURAL PATHS	
Population I	$a_{G1} = 0.424$, $b_{G1} = 0.390$; $a_{G2} = 0.424$, $b_{G2} = 0.390$
Population II	$a_{G1} = 0.000$, $b_{G1} = 0.390$; $a_{G2} = 0.141$, $b_{G2} = 0.390$
Population III	$a_{G1} = 0.000$, $b_{G1} = 0.390$; $a_{G2} = 0.424$, $b_{G2} = 0.390$
PARAMETER VALUES FOR RESIDUAL VARIANCES	
0	$\psi_{M(G1)} = 1.0$, $\psi_{Y(G1)} = 1.0$; $\psi_{M(G2)} = 1.0$, $\psi_{Y(G2)} = 1.0$
M1	$\psi_{M(G1)} = 0.5$, $\psi_{Y(G1)} = 1.0$; $\psi_{M(G2)} = 1.0$, $\psi_{Y(G2)} = 1.0$
M2	$\psi_{M(G1)} = 0.5$, $\psi_{Y(G1)} = 1.0$; $\psi_{M(G2)} = 1.5$, $\psi_{Y(G2)} = 1.0$
M3	$\psi_{M(G1)} = 0.5$, $\psi_{Y(G1)} = 1.0$; $\psi_{M(G2)} = 2.0$, $\psi_{Y(G2)} = 1.0$
Y1	$\psi_{M(G1)} = 1.0$, $\psi_{Y(G1)} = 0.5$; $\psi_{M(G2)} = 1.0$, $\psi_{Y(G2)} = 1.0$
Y2	$\psi_{M(G1)} = 1.0$, $\psi_{Y(G1)} = 0.5$; $\psi_{M(G2)} = 1.0$, $\psi_{Y(G2)} = 1.5$
Y3	$\psi_{M(G1)} = 1.0$, $\psi_{Y(G1)} = 0.5$; $\psi_{M(G2)} = 1.0$, $\psi_{Y(G2)} = 2.0$

21 populations were created by 3 (structural paths) by 7 (residual variances) combinations, e.g., Population I–0, Population I–M1, ..., Population III-Y3. The direct effects of X on Y $\hat{c}'_{G1} = \hat{c}'_{G2} = 0$ in all populations.

indirect effect in G2 ($a_{G1}b_{G1} = 0.000$; $a_{G2}b_{G2} = 0.055$); the group difference in the indirect effect was $\left(a_{G1}b_{G1} - a_{G2}b_{G2}\right) = -0.055$. In Population III, there was no indirect effect in G1; there was a large indirect effect in G2 ($a_{G1}b_{G1} = 0.000$; $a_{G2}b_{G2} = 0.165$); the group difference in the indirect effect was –0.165. The direct effect of X on Y was set to zero (i.e., $\hat{c}'_{G1} = \hat{c}'_{G2} = 0$) in all populations. It has been shown in a previous simulation study (Ryu, 2015) that the population value of the direct effect had little influence on the performance of the five methods for testing the group difference in indirect effect. With each set of the parameter values for structural paths, there were seven patterns of residual variances of M and Y. In Population -0, the residual variances of M and Y were equal between the groups in the population. In Populations -M1, -M2, and -M3, the residual variance of M was smaller in G1. In Populations -Y1, -Y2, and -Y3, the residual variance of Y was smaller in G1. Note that the effect sizes varied depending on the residual variances. The proportions of explained variance in M and Y in the 21 populations are summarized in **Table 3**.

We considered three different sample sizes for each of the 21 populations. Sample size 1: $n_{G1} = 150$; $n_{G2} = 150$. Sample size 2: $n_{G1} = 200$; $n_{G2} = 100$. Sample size 3: $n_{G1} = 100$; $n_{G2} = 200$. With Sample size 2, the residual variances were smaller in the larger group. With Sample size 3, the residual variances were smaller in the smaller group. We used Mplus 7 for data generation and estimation (Muthén and Muthén, 1998–2012). We used SAS PROC IML for resampling of the data to create bootstrap samples. We conducted 1,000 replications in each condition.

TABLE 3 | Proportion of explained variance in M and Y in population.

Population	Group 1		Group 2	
	M	Y	M	Y
I-0	0.152	0.152	0.152	0.152
I-M1	0.264	0.094	0.152	0.152
I-M2	0.264	0.094	0.107	0.204
I-M3	0.264	0.094	0.082	0.249
I-Y1	0.152	0.264	0.152	0.152
I-Y2	0.152	0.264	0.152	0.107
I-Y3	0.152	0.264	0.152	0.082
II-0	0.000	0.132	0.019	0.134
II-M1	0.000	0.071	0.019	0.134
II-M2	0.000	0.071	0.013	0.188
II-M3	0.000	0.071	0.010	0.235
II-Y1	0.000	0.233	0.019	0.134
II-Y2	0.000	0.233	0.019	0.094
II-Y3	0.000	0.233	0.019	0.072
III-0	0.000	0.132	0.152	0.152
III-M1	0.000	0.071	0.152	0.152
III-M2	0.000	0.071	0.107	0.204
III-M3	0.000	0.071	0.082	0.249
III-Y1	0.000	0.233	0.152	0.152
III-Y2	0.000	0.233	0.152	0.107
III-Y3	0.000	0.233	0.152	0.082

*See **Table 2** for population parameter values.*

We analyzed each of the generated data sets both in single-group analysis (0 = Group 1, 1 = Group 2) and in multi-group analysis to test the group difference in *a* path, the group difference in the indirect effect of X on Y via M, and the simple indirect effect in each group. We used the methods summarized in **Table 1**. We provide the sample syntax for data generation and analysis in the Appendix.

Evaluation of Methods

In order to check the data generation and estimation, we first examined the bias of the estimates. Bias was computed by (mean of estimates–true value in the population). Relative bias was computed by (bias/true value in the population) for the effects whose population values were not zero. In the single-group analysis, we compared the following estimates to their corresponding population values: individual path coefficients $\hat{a}_1$, $\hat{a}_3$, $\hat{b}$, the simple indirect effect in Group 1 $\hat{a}_1\hat{b}$, and the simple indirect effect in Group 2 $\left(\hat{a}_1 + \hat{a}_3\right)\hat{b}$. In the multi-group analysis, we compared the following estimates to their corresponding population values: individual path coefficients $\hat{a}_{G1}$, $\hat{b}_{G1}$, $\hat{a}_{G2}$, $\hat{b}_{G2}$, the simple indirect effects in each group $\hat{a}_{G1}\hat{b}_{G1}$, $\hat{a}_{G2}\hat{b}_{G2}$, and the group difference in the indirect effect $\left(\hat{a}_{G1}\hat{b}_{G1} - \hat{a}_{G2}\hat{b}_{G2}\right)$.

To evaluate the performance of the methods, we examined the rejection rates that can be interpreted as Type I error rate (when the effect was zero in population) or statistical power (when there was a non-zero effect in population) for each method. For the z test of a_3 path (z^S_{a3}), LR test (LR^M_a, LR^M_{diff}), and Wald test (W^S_{diff}, W^M_{diff}), we used $\alpha = 0.05$ criterion. For confidence intervals (95%), we computed the rejection rate by the proportion of replications in which the interval estimates did not include zero. We also examined coverage rates, width of confidence intervals, rate of left-side misses, rate of right-side misses, and ratio of left-side misses to right-side misses for interval estimates.

RESULTS

As expected, the estimates were unbiased in all populations with all sample sizes. In the single-group analysis, the bias ranged from 0.007 to −0.005, and the relative bias ranged from −0.038 to 0.007. The estimates obtained in the single-group analysis were unbiased regardless of whether the assumption of equal residual variances was satisfied. In the multi-group analysis, the bias ranged from −0.004 to 0.007, and the relative bias ranged from −0.011 to 0.051.

We present the simulation results in three sections: methods for testing the group difference in *a* path, methods for testing the group difference in the indirect effect, and methods for testing simple indirect effect in each group.

Group Difference in *a* Path

Table 4 shows the empirical Type I error rates (nominal $\alpha = 0.05$) of the methods for testing the group difference in *a* path in single-group (z^S_{a3}) and multi-group analysis (LR^M_a) in Population I.

TABLE 4 | Type I error rates of the methods for testing group difference in a path.

	Sample size					
	$n_{G1} = 150; n_{G2} = 150$		$n_{G1} = 200; n_{G2} = 100$		$n_{G1} = 100; n_{G2} = 200$	
Population	z^S_{a3}	LR^M_a	z^S_{a3}	LR^M_a	z^S_{a3}	LR^M_a
I-0	0.051	0.049	0.052	0.051	0.056	0.053
I-M1	0.055	0.052	**0.086**	0.056	0.031	0.053
I-M2	0.048	0.051	**0.113**	0.057	**0.019**	0.058
I-M3	0.048	0.047	**0.129**	0.057	**0.015**	0.056
I-Y1	0.051	0.049	0.052	0.051	0.056	0.053
I-Y2	0.051	0.049	0.052	0.051	0.056	0.053
I-Y3	0.051	0.049	0.052	0.051	0.056	0.053

The superscripts "S" and "M" indicate the single-group and multi-group approaches, respectively. W, Wald test; LR, likelihood ratio test. See ***Table 1*** *for description of each method. See* ***Table 2*** *for population parameter values. The Type I error rates that are smaller than 0.025 or greater than 0.075 are shown in bold.*

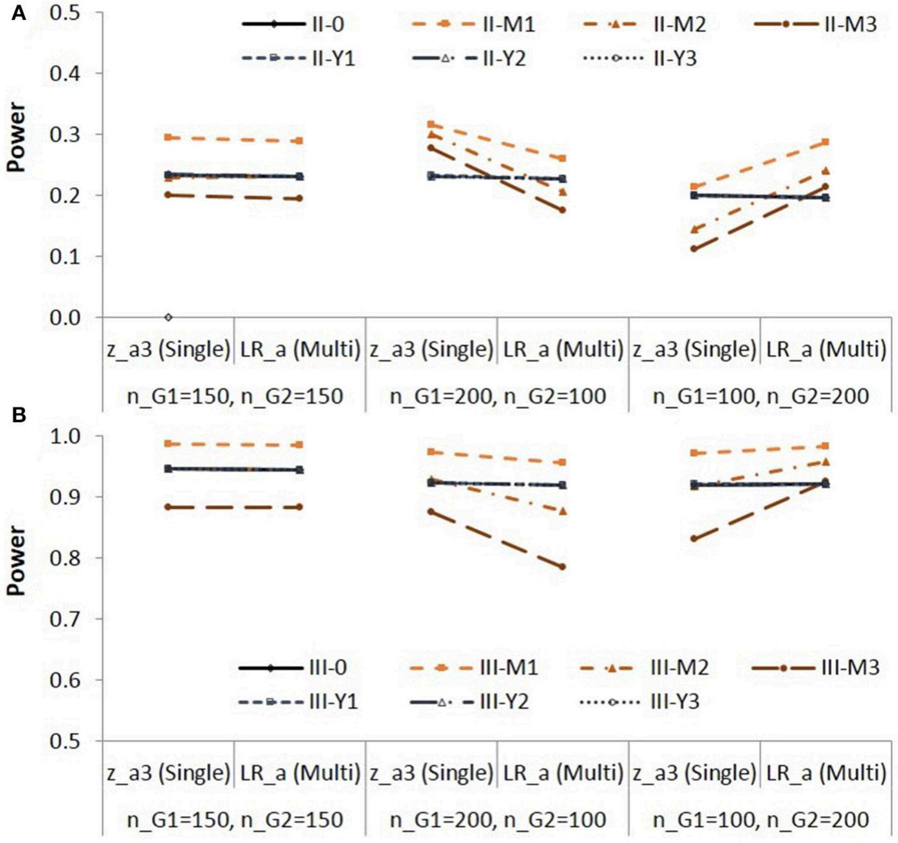


FIGURE 3 | Empirical power for testing group difference in X to M relationship (*a* path) in Population II (A) and in Population III **(B)**. See **Table 1** for description of the methods.

The Type I error rates of the LR^M_a method stayed close to the nominal level. But the z^S_{a3} method resulted in inflated Type I error rates when the residual variance of M was smaller in the group with a larger sample size (Populations I-M1 to I-M3; $n_{G1} = 200$; $n_{G2} = 100$). The z^S_{a3} method resulted in deflated Type I error rates when the residual variance of M was smaller in the group with a smaller sample size (Populations I-M2 and I-M3; $n_{G1} = 100$; $n_{G2} = 200$). Whether or not the residual variance of Y was equal between groups did not affect the Type I error rates of the z^S_{a3} method. **Figure 3** shows the empirical power of the two methods for Populations II and III.

Note that the effect sizes are different in different populations. **Figure 3** is to compare the two methods z^S_{a3} and LR^M_a in each condition. When the group sizes were equal, the power was similar for the two methods. When the residual variance of M was not equal (Populations II-M1 to II-M3, Populations III-M1 to III-M3), the z^S_{a3} method showed higher power than the LR^M_a method with the Sample size 2 ($n_{G1} = 200$; $n_{G2} = 100$); the z^S_{a3} method showed lower power than the LR^M_a method with the Sample size 3 ($n_{G1} = 100$; $n_{G2} = 200$).

Group Difference in the Indirect Effect

Type I Error Rates

Table 5 shows the empirical Type I error rates of the methods for testing the group difference in the indirect effect in Population I.

The Type I error rates for the W^S_{diff} method were higher than the nominal level when the residual variance of M was smaller in the group with a larger sample size (Populations I-M2 and I-M3; $n_{G1} = 200$; $n_{G2} = 100$); and the Type I error rates were smaller than the nominal level when the residual variance of M was smaller in the group with a smaller sample size (Populations I-M1 to I-M3; $n_{G1} = 100$; $n_{G2} = 200$). This is a similar pattern to the Type I error rates of the z^S_{a3} method in **Table 4**.

For the five methods in the multi-group analysis, the Type I error rates ranged from 0.049 to 0.068 with Sample size 1; ranged from 0.047 to 0.070 with Sample size 2; and ranged from 0.053 to 0.065 with Sample size 3. The equality of residual variances of M and Y in the population did not affect the Type I error rates of the five methods in the multi-group analysis. The Type I error rates of the BC^M_{diff} method were slightly higher than the Type I error rates of the other methods.

TABLE 5 | Type I error rates of the methods for testing group difference in the indirect effect.

Population	W^S_{diff}	LR^M_{diff}	W^M_{diff}	PC^M_{diff}	BC^M_{diff}	MC^M_{diff}
SAMPLE SIZE 1: $n_{G1} = 150$; $n_{G2} = 150$						
I-0	0.040	0.060	0.058	0.061	0.067	0.062
I-M1	0.037	0.054	0.051	0.049	0.055	0.057
I-M2	0.036	0.057	0.055	0.060	0.062	0.056
I-M3	0.039	0.058	0.053	0.066	0.065	0.063
I-Y1	0.042	0.061	0.062	0.067	0.063	0.060
I-Y2	0.040	0.066	0.065	0.059	0.068	0.065
I-Y3	0.038	0.066	0.062	0.062	0.066	0.065
SAMPLE SIZE 2: $n_{G1} = 200$; $n_{G2} = 100$						
I-0	0.044	0.050	0.054	0.056	0.060	0.051
I-M1	0.063	0.047	0.047	0.058	0.053	0.051
I-M2	**0.086**	0.053	0.055	0.055	0.061	0.053
I-M3	**0.105**	0.057	0.059	0.055	0.062	0.062
I-Y1	0.047	0.056	0.058	0.061	0.064	0.055
I-Y2	0.046	0.058	0.060	0.059	0.064	0.059
I-Y3	0.044	0.057	0.059	0.064	0.070	0.060
SAMPLE SIZE 3: $n_{G1} = 100$; $n_{G2} = 200$						
I-0	0.049	0.054	0.054	0.060	0.059	0.054
I-M1	**0.018**	0.054	0.053	0.054	0.064	0.058
I-M2	**0.011**	0.056	0.056	0.057	0.061	0.060
I-M3	**0.010**	0.055	0.057	0.057	0.058	0.055
I-Y1	0.050	0.058	0.059	0.057	0.064	0.061
I-Y2	0.047	0.060	0.056	0.058	0.064	0.062
I-Y3	0.042	0.060	0.059	0.063	0.065	0.059

*The superscripts "S" and "M" indicate the single-group and multi-group approaches, respectively. W, Wald test; LR, likelihood ratio test; PC, percentile bootstrap; BC, bias-corrected bootstrap; MC, Monte Carlo method. See **Table 1** for description of each method. The Type I error rates that are smaller than 0.025 or greater than 0.075 are shown in bold.*

Power

The empirical power for testing the group difference in the indirect effect in Populations II and III are shown in **Figure 4**.

Note that the difference in empirical power across populations (i.e., across different lines) are due to different effect sizes as shown in **Table 3**. The BC^M_{diff} method showed higher power than the other methods. The W^M_{diff} method showed lower power than the other methods in multi-group analysis. For Population III in which the group difference in the indirect effect was larger, the differences in empirical power between the methods were greater with the sample size $n_{G1} = 200$; $n_{G2} = 100$, i.e., when the indirect effect was zero in the larger group and larger in the smaller group. When the residual variance of M was not equal between groups (e.g., II-M1,..., II-M3, III-M1,..., III-M3), the W^S_{diff} method yielded higher power than the other methods with the sample size $n_{G1} = 200$; $n_{G2} = 100$. Note that the W^S_{diff} method showed inflated Type I error rates in these conditions. The W^S_{diff} method yielded lower power than the other methods with the sample size $n_{G1} = 100$; $n_{G2} = 200$. In these conditions, the Type I error rates were lower than the nominal level.

Coverage Rates, Width, and Misses

Three methods in multi-group analysis produced 95% confidence intervals for the group difference in the indirect effect: PC^M_{diff}, BC^M_{diff}, and MC^M_{diff}. The results showed similar patterns in all simulation conditions. The performance of the three confidence intervals was comparable in terms of coverage, width, and misses. The coverage rates of the PC^M_{diff} confidence intervals ranged from 0.927 to 0.951 (average = 0.939). The coverage rates of the BC^M_{diff} confidence intervals ranged from 0.923 to 0.947 (average = 0.935). The coverage rates of the MC^M_{diff} confidence intervals ranged from 0.926 to 0.949 (average = 0.934). On average, the coverage rates were slightly lower than the nominal level. The width of the confidence intervals produced by the three methods was similar to one another. The average width was 0.248 for PC^M_{diff}, 0.250 for BC^M_{diff}, and 0.246 for MC^M_{diff}.

For PC^M_{diff}, the average ratio of left-to right-side misses was 1.427, 1.927, and 1.824 in Populations I, II, and III, respectively. For BC^M_{diff}, the average ratio was 1.274, 1.521, and 1.249 in Populations I, II, and III, respectively. For MC^M_{diff}, the average ratio was 1.397, 1.783, and 1.664 in Populations I, II, and III, respectively. All three confidence intervals showed higher rates of left-side misses than right-side misses[2]. The BC^M_{diff} confidence intervals were most balanced (i.e., average ratio closer to 1).

[2] The confidence intervals were obtained for $(a_{G1}b_{G1} - a_{G2}b_{G2})$. The rates of left-side and right-side misses would be reversed if the group difference is calculated in the opposite direction.

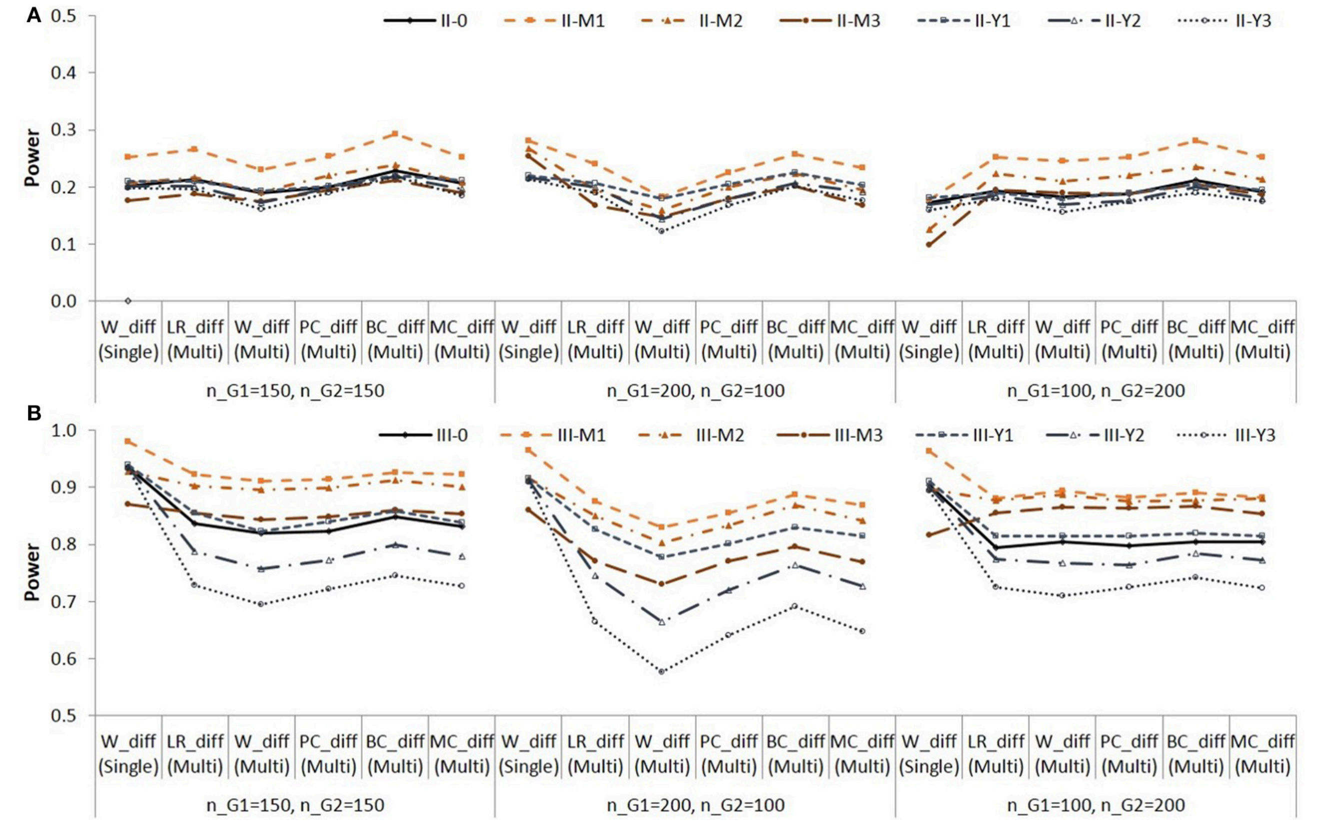


FIGURE 4 | Empirical power for testing group difference in the indirect effect in Population II (A) and in Population III **(B)**. See **Table 1** for description of the methods.

Simple Indirect Effect in Each Group

Type I Error Rates

The indirect effect was zero in Group 1 in Populations II and III. The Type I error rates for testing the simple indirect effect are shown in **Table 6**. The results were similar in Populations II and III, and the results for Population II are shown in **Table 6**.

In the single-group analysis, the Type I error rates were higher for the BC^S_{ind} method than for the PC^S_{ind} method. In the multi-group analysis, the PC^M_{ind} and MC^M_{ind} methods showed the Type I error rates that were close to the nominal level. Overall, the BC^M_{ind} method resulted in higher Type I error rates than the PC^M_{ind} and MC^M_{ind} methods. The Type I error rates of the BC^M_{ind} method were greater than 0.075 in some conditions (shown in bold).

Power

Figure 5 shows the power for testing the simple indirect effect in Group 2 in Population II, in which $a = 0.141$ and $b = 0.390$. When $a = 0.424$ and $b = 0.390$ in population (i.e., both groups in Population I, and Group 2 in Population II), the power for testing the simple indirect effects in each group was very high in all conditions.

Again, note that the difference in empirical power across populations (i.e., across different lines) are due to different effect sizes as shown in **Table 3**. The BC^S_{ind} and BC^M_{diff} methods were slightly more powerful than the other methods. The PC^S_{ind}, PC^M_{ind}, and MC^M_{ind} showed similar power.

Coverage Rates, Width, and Misses

In the single-group analysis, the coverage rates of the PC^S_{ind} confidence intervals ranged from 0.926 to 0.952 (average = 0.939). The coverage rates of the BC^S_{ind} confidence intervals ranged from 0.919 to 0.950 (average = 0.934). In the multi-group analysis, the coverage rates ranged from 0.920 to 0.962 (average = 0.937) for the PC^M_{ind} method; from 0.910 to 0.953 (average = 0.932) for the BC^M_{ind} method; from 0.919 to 0.962 (average = 0.938) for the MC^M_{ind} method. The results showed similar pattern in Populations I, II, and III. We present the coverage rates for Group 1 in Population II in **Figure 6**.

The BC^S_{ind} and BC^M_{ind} methods yielded lower coverage rates than the other methods. The PC^S_{ind}, PC^M_{ind}, and MC^M_{ind} methods showed more accurate coverage rates than the BC^S_{ind} and BC^M_{ind} methods.

On average, the confidence interval methods in the multi-group analysis resulted in wider intervals than those in the single-group analysis. The average width across all conditions was 0.147 for PC^S_{ind}, and 0.148 for BC^S_{ind}. In the multi-group analysis, the average width was 0.169 for PC^M_{ind}, 0.172 for BC^M_{ind}, and 0.168 for MC^M_{ind}.

TABLE 6 | Type I error rates for testing simple indirect effect in Group 1 in Population II.

Population	PC^{S}_{ind}	BC^{S}_{ind}	PC^{M}_{ind}	BC^{M}_{ind}	MC^{M}_{ind}
SAMPLE SIZE 1: $n_{G1} = 150; n_{G2} = 150$					
II-0	0.050	0.065	0.056	0.073	0.049
II-M1	0.048	0.068	0.049	0.075	0.043
II-M2	0.048	0.067	0.047	0.071	0.042
II-M3	0.048	0.066	0.046	0.075	0.040
II-Y1	0.050	0.063	0.053	0.063	0.054
II-Y2	0.050	0.065	0.052	0.064	0.051
II-Y3	0.050	0.067	0.052	0.061	0.046
SAMPLE SIZE 2: $n_{G1} = 200; n_{G2} = 100$					
II-0	0.060	0.074	0.062	**0.082**	0.059
II-M1	0.058	0.074	0.061	**0.090**	0.059
II-M2	0.058	0.071	0.062	**0.085**	0.058
II-M3	0.058	0.072	0.055	**0.082**	0.061
II-Y1	0.060	0.066	0.062	0.069	0.061
II-Y2	0.060	0.071	0.063	0.071	0.059
II-Y3	0.060	0.071	0.065	**0.077**	0.061
SAMPLE SIZE 3: $n_{G1} = 100; n_{G2} = 200$					
II-0	0.051	0.069	0.053	**0.082**	0.056
II-M1	0.051	0.071	0.039	**0.077**	0.038
II-M2	0.051	0.067	0.040	**0.077**	0.045
II-M3	0.051	0.064	0.042	0.072	0.038
II-Y1	0.051	0.067	0.058	**0.076**	0.059
II-Y2	0.051	0.072	0.061	**0.078**	0.063
II-Y3	0.051	0.072	0.065	**0.077**	0.056

The superscripts "S" and "M" indicate the single-group and multi-group approaches, respectively. PC, percentile bootstrap; BC, bias-corrected bootstrap; MC, Monte Carlo method. See ***Table 1*** *for description of each method. Type I error rates that are smaller than 0.025 or greater than 0.075 are shown in bold.*

Table 7 shows the average ratio of left- to right-side misses of confidence intervals methods for simple indirect effects.

The confidence intervals showed higher rates of right-side misses for the simple indirect effects whose population values were positive, except BC^{M}_{ind} in Population I. The confidence intervals showed higher rates of left-side misses for simple indirect effects whose population values were zero. Both in the single-group and multi-group analysis, the bias-corrected confidence intervals, BC^{S}_{ind} and BC^{M}_{ind}, were most balanced (i.e., average ratio closer to 1).

EMPIRICAL EXAMPLE

We illustrate the methods using empirical data from PISA 2003 database (Programme for International Student Assessment, Organisation for Economic Co-operation Development, 2004, 2005). We adopted a conceptual model in Yeung (2007). We compared the indirect effect of teachers' emotional support on math interest via math self-concept in Australia (AUS; $N =$ 1,2551) and Austria (AUT; $N = 4{,}597$). The estimated multi-group and single-group structural equation models are shown in **Figure 7**. We applied the methods for (i) comparing the a path between groups, (ii) comparing the indirect effect between groups, (iii) testing simple indirect effect in each group. In the multi-group model (**Figure 7A**), with the b path (Math self-concept → Math interest) and c' path (Emotional support → Math interest) set equal between groups, $\chi^2\,(2) = 0.464$, $p =$ 0.793, CFI = 1.000, RMESA = 0.000, SRMR = 0.003. We kept the equality constrains on b and c' paths in the multi-group model so that the specification of the fixed effects is equivalent to the single-group model. In the single-group model (**Figure 7B**), we created a group variable to represent the two countries that 0 = Australia (AUS) and 1 = Austria (AUT). The results are summarized in **Table 8**. In the multi-group model, the residual variances were slightly smaller in AUS whose sample size was larger. This is similar to Sample size 2 ($n_{G1} = 200$; $n_{G2} = 100$) condition in the simulation. In **Table 8**, LR^{M}_{a} was slightly more conservative than z^{S}_{a3} in testing the group difference in a path; LR^{M}_{diff} and W^{M}_{diff} were slightly more conservative than W^{S}_{diff} in testing the group difference in the indirect effect. For the difference in the indirect effect, PC^{M}_{diff}, BC^{M}_{diff}, and MC^{M}_{diff} yielded in comparable results. For the simple indirect effect, PC^{S}_{ind}, BC^{S}_{ind}, PC^{M}_{ind}, BC^{M}_{ind}, and MC^{M}_{ind} resulted in comparable interval estimates.

SUMMARY AND DISCUSSION

When the research question involves comparing indirect effects between distinctive groups, researchers can choose single-group or multi-group analysis approach in SEM framework to incorporating the group membership as a categorical moderator. In this article, we evaluated statistical methods for (i) comparing a structural path (in our example, a path or X → M relationship) between groups, (ii) comparing the indirect effect between groups, and (iii) testing simple indirect effect in each group. We continue to use the abbreviated names of each method to summarize and discuss the results (See **Table 1**).

The key findings in the simulation study are:

(1) In the single-group analysis, the z^{S}_{a3} and W^{S}_{diff} methods may result in invalid statistical inferences when the assumption of equal variances is neglected.
(2) However, the performance of bootstrapping confidence intervals is robust even when the bootstrap estimates are obtained in the single-group model.
(3) The bias-corrected bootstrap confidence intervals are slightly more powerful than the percentile bootstrap and Monte Carlo confidence intervals, but at the cost of higher Type I error rate, and;
(4) For comparing an indirect effect between groups, the likelihood ratio test in the multi-group analysis is as powerful as the other methods with the Type I error rate staying close to the desired level.

For testing the group difference in the a path, the assumption of equal variances was critical for the z^{S}_{a3} method, but not for the LR^{M}_{a} method in the multi-group analysis. When the assumption was not satisfied, the z^{S}_{a3} method showed inaccurate Type I error rates, as expected. The Type I error rates were

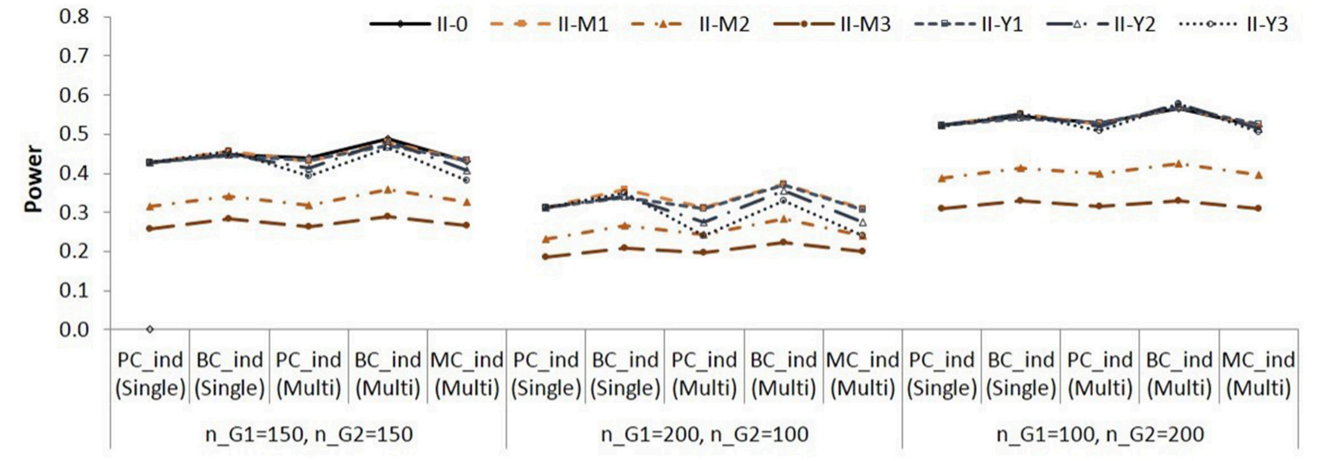


FIGURE 5 | Empirical power for testing simple indirect effect in Group 2 in Population II. See **Table 1** for description of the methods.

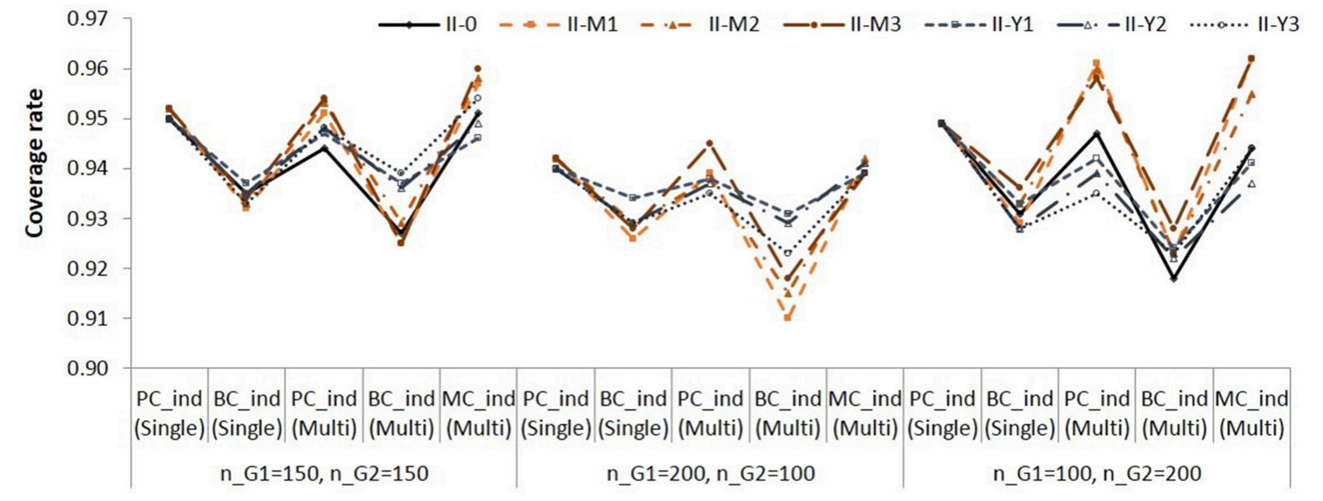


FIGURE 6 | Coverage rates of 95% confidence intervals for the simple indirect effects in Group 1 in Population II. See **Table 1** for description of the methods.

TABLE 7 | Average ratio of left-to-right misses of confidence intervals methods for simple indirect effects.

	Population I		Population II		Population III	
	Group1[a]	Group2[a]	Group1[b]	Group2[a]	Group1[b]	Group2[a]
PC_{ind}^{S}	0.627	0.486	1.674	0.606	1.674	0.494
BC_{ind}^{S}	0.969	0.791	1.467	0.770	1.476	0.790
PC_{ind}^{M}	0.613	0.474	1.670	0.493	1.644	0.491
BC_{ind}^{M}	1.025	0.851	1.491	0.660	1.486	0.886
MC_{ind}^{M}	0.604	0.506	1.626	0.488	1.638	0.499

The superscripts "S" and "M" indicate the single-group and multi-group approaches, respectively. PC, percentile bootstrap; BC, bias-corrected bootstrap; MC, Monte Carlo method. See ***Table 1*** *for description of each method.* [a]*The simple indirect effect was positive in population.* [b]*The simple indirect effect was zero in population.*

inflated when the variance was larger in the smaller group, and deflated when the variance was larger in the larger group.

For testing the simple indirect effect in each group, the bootstrap confidence intervals in the single-group analysis (PC_{ind}^{S}, BC_{ind}^{S}) were not affected by the violation of the equal

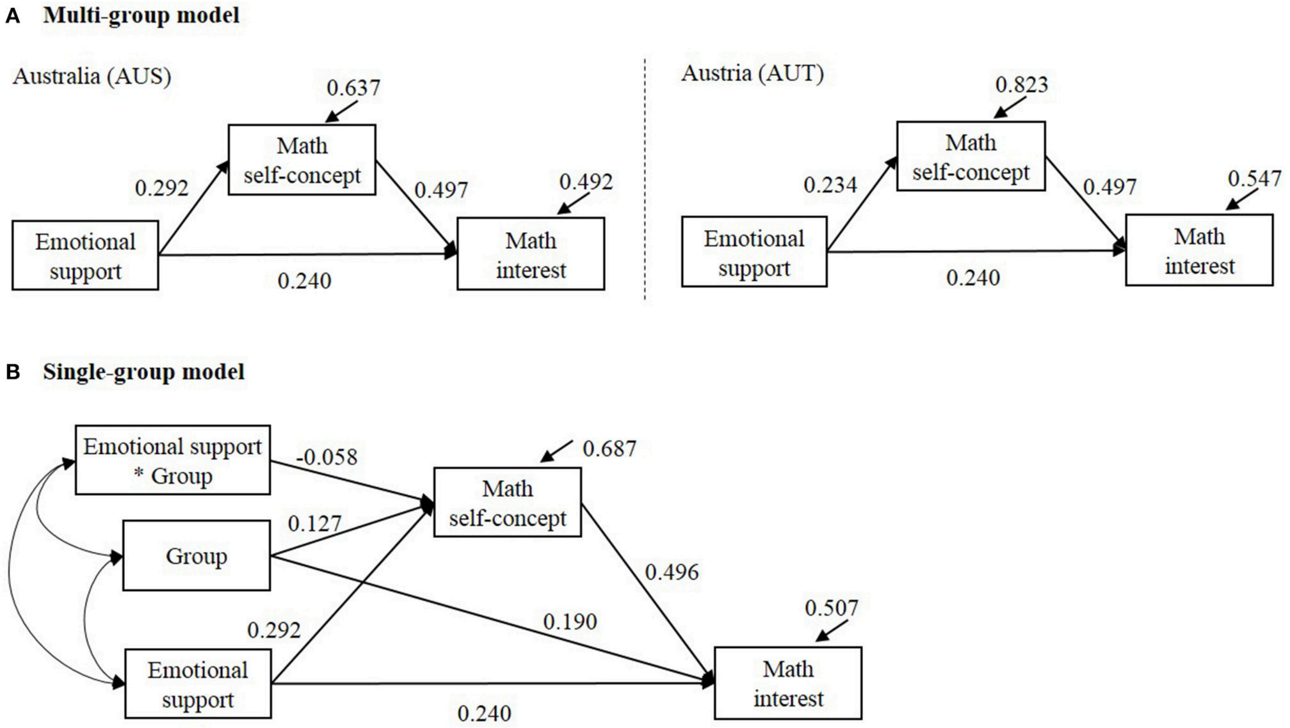


FIGURE 7 | Estimated multi-group and single-level structural equation models. In **(A)** Multi-group model, the path coefficient "Math self-concept → Math interest" was set equal between groups; the path coefficient "Emotional support → Math interest" was set equal between groups. The estimate of the indirect effect was 0.292*0.497 = 0.145 in Australia (AUS) and 0.234*0.497 = 0.116 in Austria (AUT). In **(B)** single-group model, group was coded 0 = AUS and 1 = AUT. The estimated indirect effect was 0.292*0.496 = 0.145 for AUS and (0.292–0.058)*0.496 = 0.116 for AUT. The estimated group difference in the indirect effect was 0.116–0.145 = –0.029.

variances assumption. The PC^S_{ind} and BC^S_{ind} confidence intervals were obtained based on the set of 1,000 estimates in bootstrap samples. As shown in the simulation results, the estimates in the single-group analysis model were unbiased regardless of whether the assumption of equal variances is satisfied. So the empirical sampling distribution of the indirect effect is expected to be comparable with or without the assumption of equal variances satisfied. Therefore, the bootstrap confidence intervals obtained from the empirical sampling distribution were not affected by the assumption.

In the multi-group analysis, all methods did not show differences in their performance depending on whether or not the equal variances assumption is satisfied. These results were expected, because the variances were estimated in each group separately in the multi-group model.

In both single-group and multi-group approaches, the bias-corrected bootstrap methods (BC^S_{ind}, BC^M_{ind}, BC^M_{diff}) tended to show slightly higher Type I error rates, higher statistical power, and lower coverage rates than the percentile bootstrap methods (PC^S_{ind}, PC^M_{ind}, PC^M_{diff}). This pattern of results is consistent with what has been found in previous studies (e.g., Preacher et al., 2007; Preacher and Selig, 2012; Ryu, 2015). The Monte Carlo methods (MC^M_{ind}, MC^M_{diff}) performed similarly to the percentile bootstrap methods. The Type I error rates and the coverage rates of the confidence intervals were close to the desired level in all conditions. The empirical power was slightly lower than the bias-corrected bootstrap methods, but not by much. The largest difference in power was 0.091.

For the interval estimates of the group difference in the indirect effect, the average widths were comparable for all three methods in the multi-group analysis (PC^M_{diff}, BC^M_{diff}, MC^M_{diff}). For the interval estimates of the simple indirect effects, the two methods in the single-group analysis (PC^S_{ind}, BC^S_{ind}) showed similar average widths, and the three methods in the multi-group analysis (PC^M_{ind}, BC^M_{ind}, MC^M_{ind}) showed similar average widths. The multi-group methods resulted in wider interval estimates of the simple indirect effects than the single-group methods.

The confidence intervals for the simple indirect effects were unbalanced with higher rate of left-side misses when the simple indirect effect was zero in population, and unbalanced with higher rate of right-side misses when there was a positive simple indirect effect in population. For both the group difference in the indirect effect and the simple indirect effects, the bias-corrected bootstrapping methods (BC^M_{diff}, BC^S_{ind}, BC^M_{ind}) were most balanced in terms of the ratio of left- and right-side misses.

In the multi-group analysis, the likelihood ratio test (LR^M_{diff}) and the Wald test (W^M_{diff}) performed well in terms of Type I

TABLE 8 | Empirical example results.

Method	Result
GROUP DIFFERENCE IN *a* PATH	
z^{S}_{a3}	$\hat{a}_3 = -0.058$, standard error = 0.020, $p = 0.005$
LR^{M}_{a}	LR statistic = 7.122, df = 1, $p = 0.0076$
GROUP DIFFERENCE IN THE INDIRECT EFFECT	
W^{S}_{diff}	Wald statistic = 7.903, df = 1, $p = 0.0049$
LR^{M}_{diff}	LR statistic = LR statistic = 7.122, df = 1, $p = 0.0076$
W^{M}_{diff}	Wald statistic = 7.115, df = 1, $p = 0.0076$
PC^{M}_{diff}	95% confidence intervals = (–0.057, –0.001)
BC^{M}_{diff}	95% confidence intervals = (–0.057, –0.001)
MC^{M}_{diff}	95% confidence intervals = (–0.051, –0.008)
SIMPLE INDIRECT EFFECT IN EACH GROUP	
PC^{S}_{ind}	95% confidence intervals = (0.128, 0.161) in AUS; (0.093, 0.139) in AUT
BC^{S}_{ind}	95% confidence intervals = (0.129, 0.161) in AUS; (0.093, 0.140) in AUT
PC^{M}_{ind}	95% confidence intervals = (0.128, 0.162) in AUS; (0.093, 0.139) in AUT
BC^{M}_{ind}	95% confidence intervals = (0.130, 0.163) in AUS; (0.092, 0.139) in AUT
MC^{M}_{ind}	95% confidence intervals = (0.134, 0.157) in AUS; (0.098, 0.135) in AUT

*See **Table 1** for description of each method.*

error rates. But the W^{M}_{diff} method showed lower power than the LR^{M}_{diff} and the confidence intervals methods for testing the group difference in the indirect effect. The empirical power of the LR^{M}_{diff} method was comparable to the power of PC^{M}_{diff} and MC^{M}_{diff}. These results are consistent with those found in a previous study (Ryu, 2015). In the single-group analysis, the performance of the Wald test (W^{S}_{diff}) for testing the group difference in the indirect effect was affected by the violation of the equal variance assumption, particularly with unequal group sizes. The Type I error rates were higher than the desired level when the variance was larger in the smaller group. The Type I error rates were smaller than the nominal level when the variance was larger in the larger group.

In many cases, studies are conducted to address questions on means (unconditional or conditional) and relationships between variables, and the variance estimates are often neglected. It is important for researchers to pay attention to variance estimates, even when they are not of key interest. When the research question involves moderation effect by a distinctive group membership, it is recommended that the variance parameters are examined first with no restriction that the variances are equal in all groups. When it is reasonable to assume that the variances are equal, researchers may choose to adopt single-group or multi-group analysis approach. When it is not reasonable to assume equal variances, multi-group analysis is recommended. The single-group analysis resulted in unbiased parameter estimates even with the assumption violated. But some methods for statistical inference were affected by the violation of the assumption. If single-group analysis is adopted, statistical methods must be chosen with careful consideration.

Multi-group analysis approach has advantages over single-group approach in incorporating a categorical moderator in the model. First, the multi-group approach does not depend on the assumption of equal variances, and so the parameter estimates and statistical inferences are not affected by the assumption satisfied or violated. Second, it is less complicated to specify and test the group difference in more than one indirect effect. For example, suppose that a mediation model is hypothesized in which three indirect effects are specified between one independent variable (X), three mediating variables (M1, M2, and M3), and one dependent variable (Y). In order to specify a model that allows the three indirect effects to differ between groups, the single-group approach requires at least three additional product terms to represent the interaction with the group membership. The number of required product terms can increase if there are more than two levels of the categorical moderator, or if both the relationship between X and the mediators and the relationship between the mediators and Y differ between groups. In the multi-group analysis, however, the group differences can be specified and tested without increasing the number of variables in the model.

In conclusion, when the data are from more than one distinctive group, we recommend that researchers first examine parameter estimates (including variance parameters) in each group with no restriction before choosing to adopt single-group analysis. For testing the group difference in the indirect effect in multi-group analysis, the likelihood ratio test is more powerful than Wald test, with Type I error rate close to the desired level. For confidence intervals of the group difference in the indirect effect, bias-corrected bootstrap confidence intervals were more powerful and more balanced than the percentile bootstrap and Monte Carlo confidence intervals, but at the cost of higher Type I error rates and lower coverage rates. For the simple indirect effect in each group, bias-corrected bootstrap confidence intervals were more powerful than the percentile bootstrap and Monte Carlo confidence intervals, but again the Type I error rates were higher with bias-corrected bootstrap confidence intervals. Taken together, we recommend the likelihood ratio test along with the percentile or Monte Carlo interval estimates for the group difference in the indirect effect. We recommend the percentile or Monte Carlo interval estimates for the simple indirect effect.

AUTHOR CONTRIBUTIONS

ER and JC designed the study. ER conducted the simulation study and took a leading role in writing the manuscript. JC conducted a part of the simulation and participated in writing.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpsyg.2017.00747/full#supplementary-material

REFERENCES

Aguinis, H., and Pierce, C. A. (1998). Heterogeneity of error variance and the assessment of moderating effects of categorical variables: a conceptual review. *Organ. Res. Methods* 1, 296–314. doi: 10.1177/109442819813002

Alwin, D. F., and Hauser, R. M. (1975). The decomposition of effects in path analysis. *Am. Sociol. Rev.* 40, 35–47. doi: 10. 2307/2094445

Bentler, P. M., and Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis covariance structures. *Psychol. Bull.* 88, 588–606.

Bollen, K. A. (1987). "Total, direct, and indirect effects in structural equation models," in *Sociological Methodology,* ed C. C. Clogg (Washington, DC: American Sociological Association), 37–69.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. New York, NY: Wiley.

Bollen, K. A., and Stine, R. (1990). Direct and indirect effects: classical and bootstrap estimates of variability. *Sociol. Methodol.* 20, 115–140. doi: 10.2307/271084

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification. *Ann. Math. Stat.* 25, 290–302. doi: 10.1214/aoms/1177728786

Dretzke, B. J., Levin, J. R., and Serlin, R. C. (1982). Testing for regression homogeneity under variance heterogeneity. *Psychol. Bull.* 91, 376–383. doi: 10.1037/0033-2909.91.2.376

Edwards, J. R., and Lambert, L. S. (2007). Methods for integrating moderation and mediation: a general analytical framework using moderated path analysis. *Psychol. Methods* 12, 1–22. doi: 10.1037/1082-989X.12.1.1

Gelfand, M. J., Brett, J., Gunia, B. C., Imai, L., Huang, T.-J., and Hsu, B.-F. (2013). Toward a culture-by-context perspective on negotiation: negotiating teams in the United States and Taiwan. *J. Appl. Psychol.* 98, 504–513. doi: 10.1037/a0031908

Glass, G. V., Peckham, P. D., and Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance. *Rev. Educ. Res.* 42, 237–288. doi: 10.3102/00346543042003237

Hayes, A. F. (2015). An index and test of linear moderated mediation. *Multivariate Behav. Res.* 50, 1–22. doi: 10.1080/00273171.2014.962683

Levant, R. F., Parent, M. C., McCurdy, E. R., and Bradstreet, T. C. (2015). Moderated mediation of the relationships between masculinity ideology, outcome expectations, and energy drink use. *Health Psychol.* 34, 1100–1106. doi: 10.1037/hea0000214

MacCallum, R. C., Zhang, S., Preacher, K. J., and Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychol. Methods* 7, 19–40.

MacKinnon, D. P., Lockwood, C. M., and Williams, J. M. (2004). Confidence limits for the indirect effect: distribution of the product and resampling methods. *Multivariate Behav. Res.* 39, 99–128. doi: 10.1207/s15327906mbr3901_4

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., and Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychol. Methods* 7, 83–104. doi: 10.1037//1082-989X.7.1.83

Maxwell, S. E., and Delaney, H. D. (1993). Bivariate median splits and spurious statistical significance. *Psychol. Bull.* 113, 181–190. doi: 10.1037/0033-2909.113.1.181

Muthén, L. K., and Muthén, B. O. (1998–2012). *Mplus User's Guide*, *7th Edn*. Los Angeles, CA: Muthén & Muthén.

Organisation for Economic Co-operation and Development (2004). *Learning for Tomorrow's World: First Results from PISA 2003*. Paris: Organisation for Economic Co-operation and Development.

Organisation for Economic Co-operation and Development (2005). *PISA 2003 Technical Report*. Paris: Organisation for Economic Co-operation and Development.

Preacher, K. J., and Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behav. Res. Methods Instrum. Comput.* 36, 717–731. doi: 10.3758/BF03206553

Preacher, K. J., and Selig, J. P. (2012). Advantages of Monte Carlo confidence intervals for indirect effects. *Commun. Methods Meas.* 6, 77–98. doi: 10.1080/19312458.2012.679848

Preacher, K. J., Rucker, D. D., and Hayes, A. F. (2007). Addressing moderated mediation hypotheses: theory, methods, and prescriptions. *Multivariate Behav. Res.* 42, 185–227. doi: 10.1080/00273170701341316

Rucker, D. D., McShane, B. B., and Preacher, K. J. (2015). A researcher's guide to regression, discretization, and median splits of continuous variables. *J. Consum. Psychol.* 25, 666–678. doi: 10.1016/j.jcps.2015.04.004

Ryu, E. (2015). Multi-group analysis approach to testing group different in indirect effects. *Behav. Res. Methods* 47, 484–493. doi: 10.3758/s13428-014-0485-8

Schnitzspahn, K. M., Thorley, C., Phillips, L., Voigt, B., Threadgold, E., Hammond, E. R., et al. (2014). Mood impairs time-based prospective memory in young but not older adults: the mediating role of attentional control. *Psychol. Aging* 29, 264–270. doi: 10.1037/a0036389

Shrout, P. E., and Bolger, N. (2002). Mediation in experimental and nonexperimental studies: new procedures and recommendations. *Psychol. Methods* 7, 422–445. doi: 10.1037/1082-989X.7.4.422

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Trans. Am. Math. Soc.* 54, 426–482. doi: 10.2307/1990256

Yeung, K. (2007). *Perception of Teacher Emotional Support and Parental Education Level: The Impacts on Students' Math Performance.* Unpublished doctoral dissertation, University of Leicester.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 13 March 2018
doi: 10.3389/fpsyg.2018.00251

Using iMCFA to Perform the CFA, Multilevel CFA, and Maximum Model for Analyzing Complex Survey Data

Jiun-Yu Wu[1]*, *Yuan-Hsuan Lee*[2] *and John J. H. Lin*[3]*

[1] Institute of Education & Center for Teacher Education, National Chiao Tung University, Hsinchu, Taiwan, [2] Department of Education and Learning Technology, National Tsing Hua University, Hsinchu, Taiwan, [3] Office of Institutional Research, National Central University, Taoyuan, Taiwan

To construct CFA, MCFA, and maximum MCFA with LISREL v.8 and below, we provide iMCFA (integrated Multilevel Confirmatory Analysis) to examine the potential multilevel factorial structure in the complex survey data. Modeling multilevel structure for complex survey data is complicated because building a multilevel model is not an infallible statistical strategy unless the hypothesized model is close to the real data structure. Methodologists have suggested using different modeling techniques to investigate potential multilevel structure of survey data. Using iMCFA, researchers can visually set the between- and within-level factorial structure to fit MCFA, CFA and/or MAX MCFA models for complex survey data. iMCFA can then yield between- and within-level variance-covariance matrices, calculate intraclass correlations, perform the analyses and generate the outputs for respective models. The summary of the analytical outputs from LISREL is gathered and tabulated for further model comparison and interpretation. iMCFA also provides LISREL syntax of different models for researchers' future use. An empirical and a simulated multilevel dataset with complex and simple structures in the within or between level was used to illustrate the usability and the effectiveness of the iMCFA procedure on analyzing complex survey data. The analytic results of iMCFA using Muthen's limited information estimator were compared with those of Mplus using Full Information Maximum Likelihood regarding the effectiveness of different estimation methods.

Keywords: multilevel structural equation modeling, confirmatory factor analysis, complex survey data, Lisrel, Mplus, maximum model

OPEN ACCESS

Edited by:
Ehri Ryu,
Boston College, United States

Reviewed by:
Salvador Chacón-Moscoso,
Universidad de Sevilla, Spain
Yu Liu,
University of Houston, United States
Ronny Scherer,
Centre for Educational Measurement,
Faculty of Educational Sciences,
University of Oslo, Norway

****Correspondence:***
Jiun-Yu Wu
jiunyu.rms@gmail.com
John J. H. Lin
john.jrhunglin@gmail.com

Specialty section:
This article was submitted to
Quantitative Psychology and
Measurement,
a section of the journal
Frontiers in Psychology

Received: *13 July 2017*
Accepted: *15 February 2018*
Published: *13 March 2018*

Citation:
Wu J-Y, Lee Y-H and Lin JJH (2018)
Using iMCFA to Perform the CFA,
Multilevel CFA, and Maximum Model
for Analyzing Complex Survey Data.
Front. Psychol. 9:251.
doi: 10.3389/fpsyg.2018.00251

INTRODUCTION

Confirmatory Factor Analysis (CFA) has been widely utilized to examine the factorial structure of measures/scales in behavioral, sociological, educational, and organizational fields (Thompson, 2004; Kaplan, 2008; Kline, 2016). Researchers utilize CFAs to examine the reliability and validity of the underlying structure of test items and the theoretical constructs (Raykov, 2004; Raykov and Marcoulides, 2006; Geldhof et al., 2013). A fundamental assumption of the CFA analysis is that the responses from participants are independently and identically distributed (Bollen, 1989; Kaplan, 2008; Kline, 2016). However, the independence assumption can hardly be met for the survey dataset in the empirical studies. For instance, in educational and organizational research, we might utilize the complex survey sampling strategy (e.g., multistage sampling, cluster sampling, etc.) to collect the responses of an individual or lower sampling unit, which are nested within between-level clusters/groups (Stapleton, 2006, 2008; e.g., Wu et al., 2014; Wu, 2017). Within this context, participants in the same group with the same cluster information might yield more homogenous responses than those in different groups (Bovaird, 2007). Using the CFA without considering the dependent/multilevel structure in the complex survey data will result in biased parameter estimates and erroneous standard error estimates as well as inconsistent statistical inferences of the analytic results (Muthén and Satorra, 1995; Stapleton, 2008; Wu and Kwok, 2012; Wu et al., 2017).

In order to examine the multilevel factorial structure of complex survey data, various CFA techniques have been proposed, such as the model-based approach (Multilevel CFA, MCFA, e.g., Muthén, 1991; Hox, 1993; Mehta and Neale, 2005) and the maximum model in CFA (MAX CFA, e.g., Ryu and West, 2009; Wu and Kwok, 2012). The MCFA builds up a hierarchical statistical model corresponding to the multilevel structure of the complex survey data, so that the within-cluster and between-cluster model parameters can be separately and freely estimated (Muthén and Satorra, 1989). MAX MCFA is a special case of MCFA (Ryu and West, 2009; Wu and Kwok, 2012), which is usually considered as a partially saturated model during the process of building up a valid MCFA model (Hox, 2010). When using the maximum modeling strategy, researchers build up a MCFA model with a saturated between-level and a hypothesized within-level model. By doing so, all the unique elements of the variance-covariance matrix in the between-level will be estimated with the consumption of all the available degrees of freedom. Therefore, the saturated and just-identified between-level model contributes nothing to the fitting function (Hox, 2010), which allows us to diagnose the misspecification of the within-level model with the level-specific model-fit information (Ryu and West, 2009). These two approaches have been shown to yield consistent parameter estimates and statistical inferences as the population multilevel model (Wu and Kwok, 2012). However, modeling the multilevel structure of complex survey data is more complicated and requires more advanced statistical techniques and specific computer software.

The purpose of this study is threefold. First, the study intends to provide an integrated software for flexible multilevel modeling with Lisrel v.8 and below. Second, we investigate the performance of CFA, MCFA, and MAX MCFA in analyzing multilevel data with a complex within and simple between structure[1] as well as a complex between and simple within structure. Third, we compare analysis results for the three modeling techniques using Muthen's limited information estimator (MUML in iMCFA) and Full Information Maximum Likelihood (FIML in Mplus). A review of literature on different ways of multilevel model construction and constraints of the current SEM software was provided, followed by the demonstration of iMCFA (i.e., the integrated Multilevel Confirmatory Analysis program).

Multilevel Model Construction

Researchers have constructed MCFA models in two major approaches. For the first approach, they separated the level-varying covariance components from total covariance structures and used the level-specific covariance component to build the specific-level models (Muthén, 1994; Yuan and Bentler, 2007). For the second approach, they used the maximum model (or the unrestricted/saturated model) as the baseline to construct the between-level model with theoretical evidence (Yuan and Bentler, 2003; Stapleton, 2008; Hox, 2010).

The basic idea of MCFA is to decompose the total variance-covariance matrix, $\boldsymbol{\Sigma}_T$, into between-level variance-covariance (V-C) matrix, $\boldsymbol{\Sigma}_B$, and within-level V-C matrix, $\boldsymbol{\Sigma}_W$. Assuming $\boldsymbol{y}_{\mathrm{gi}}$ is the observed variables for participant i within cluster g, the total V-C matrix $\boldsymbol{\Sigma}_T = Var\left[\boldsymbol{y}_{\mathrm{gi}}\right]$. The corresponding between- and within-level V-C components will be orthogonal and additive (Searle et al., 1992; Muthén, 1994). Same score decomposition can be performed for the observed complex survey sample data, and the resulted sample V-C matrix can be shown as,

$$\mathbf{S_T} = \mathbf{S_B}+\mathbf{S_W}$$

where $\mathbf{S_B}$ and $\mathbf{S_W}$ are the level-varying V-C estimators to their population counterparts, $\boldsymbol{\Sigma}_\mathbf{B}$ and $\boldsymbol{\Sigma}_\mathbf{W}$, respectively (Muthén, 1994; Hox, 2002; Hox and Maas, 2004; Heck and Thomas, 2008). With the variance-covariance matrix decomposition, Muthén (1989, 1990) presented an a partial Maximum likelihood estimation method, also named MUML (Muthén's limited information estimator). In MUML, two variance-covariance matrices of different levels are constructed as

$$\mathbf{S_T} = \mathbf{S_{B,MUML}}+\mathbf{S_{PW,MUML}} \tag{1}$$

Consider a multilevel dataset with the sample size of N, i.e., on average N_g participants nested within respective G groups. The above three variance-covariance matrices are defined as

$$\begin{aligned}
\mathbf{S_T} &= \frac{1}{N-1}\sum_{g=1}^{G}\sum_{i=1}^{N_g}\left(\mathbf{y}_{gi}-\bar{\mathbf{y}}\right)\left(\mathbf{y}_{gi}-\bar{\mathbf{y}}\right)' \\
\mathbf{S_{PW,MUML}} &= \frac{1}{N-G}\sum_{g=1}^{G}\sum_{i=1}^{N_g}\left(\mathbf{y}_{gi}-\bar{\mathbf{y}}_g\right)\left(\mathbf{y}_{gi}-\bar{\mathbf{y}}_g\right)' \\
\mathbf{S_{B,MUML}} &= \frac{1}{G-1}\sum_{g=1}^{G}\left(\bar{\mathbf{y}}_g-\bar{\mathbf{y}}\right)\left(\bar{\mathbf{y}}_g-\bar{\mathbf{y}}\right)'
\end{aligned} \tag{2}$$

where the grand mean $\bar{\mathbf{y}} = \frac{1}{N}\sum_{g=1}^{G}\sum_{i=1}^{N_g}\mathbf{y}_{gi}$ and group mean of g^{th} group $\bar{\mathbf{y}}_g = \frac{1}{N_g}\sum_{i=1}^{N_g}\mathbf{y}_{gi}$.

In Equation (2), Muthén showed that the pooled within-level observed variance-covariance matrix $\mathbf{S_{PW,MUML}}$ is the consistent and unbiased estimator to $\boldsymbol{\Sigma}_\mathbf{W}$, and the scaled between-level observed variance-covariance matrix $\mathbf{S_{B,MUML}}$ is the consistent and unbiased estimator to $\boldsymbol{\Sigma}_\mathbf{W} + c\boldsymbol{\Sigma}_\mathbf{B}$, where $c = (N(G-1))^{-1}\left(N^2 - \sum_{g}^{G} n_g^2\right)$ is close to the averaged group size.

[1] In this study, we refer simple structure CFA as the model with one factor, and complex structure CFA as the model with more than one factors.

In a balance design (i.e., all between-level units have the same group size), MUML is the same as the original unbiased ML estimator. But in an unbalance design, MUML is the simplified version of quasi-ML estimation method (Varin and Vidoni, 2005) and only uses a common group size, *c*, as the weighting scalar of the between-level variance component in the likelihood function, that is,

$$\begin{aligned} F_{MUML}\left(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}}\right) &= F_{MUML}\left(\mathbf{S}, \hat{\boldsymbol{\Sigma}}\right) = G\left\{\ln\left|\hat{\boldsymbol{\Sigma}}_{\mathbf{W}} + c\hat{\boldsymbol{\Sigma}}_{\mathbf{B}}\right|\right. \\ &\quad \left. + tr\left(\left(\hat{\boldsymbol{\Sigma}}_{\mathbf{W}} + c\hat{\boldsymbol{\Sigma}}_{\mathbf{B}}\right)^{-1}\mathbf{S}_{\mathbf{B}}\right) - \ln\left|\mathbf{S}_{\mathbf{B}}\right| - p\right\} \\ &\quad + (N-G)\left\{\ln\left|\hat{\boldsymbol{\Sigma}}_{\mathbf{W}}\right| + tr\left(\hat{\boldsymbol{\Sigma}}_{\mathbf{W}}^{-1}\mathbf{S}_{\mathbf{PW}}\right)\right. \\ &\quad \left. - \ln\left|\mathbf{S}_{\mathbf{PW}}\right| - p\right\} \qquad (3) \end{aligned}$$

MUML is also called as limited information or quasi-maximum likelihood estimation because it assumes that all groups have equal group size, even though they may not. Researchers can use the MUML in Mplus (Muthén and Muthén, 1998-2017) with the routine of "ESTIMATOR=MUML."

Due to the limitation of conducting MCFA analyses in LISREL v.8 and below, we can decompose the between- and within-level variance-covariance structures shown in Equation (2). One nice feature about MUML is that researchers can use the multi-group analysis routine provided in various SEM programs to conduct the multilevel CFA analysis. Researchers need to separate the original data into two groups: the between-level group with between-level V-C matrix $\mathbf{S}_{\mathbf{B,MUML}}$ and group number *G*, and the within-level group with within-level V-C matrix $\mathbf{S}_{\mathbf{PW},\mathbf{MUML}}$ with sample size *N-G*. The multilevel data can then be analyzed with the multi-group routine. The detailed steps of this process is provided in Heck and Thomas (2008) and Muthén (1994). Compared with Full-information Maximum Likelihood estimator (FIML, Arbuckle, 1996; Mehta and Neale, 2005), MUML is simpler in computing the parameter estimates while FIML is computationally heavier as the size of sub-groups increases. Muthén and Satorra (1995) concluded that MUML generally performs equally well as FIML in various conditions; however, Hox and Maas (2004) showed FIML has more accurate parameter estimates than MUML does. We will check the analytical result consistency between iMCFA using MUML and Mplus using FIML with unbalanced- and balanced-design[2] samples in the provided scenarios.

Multilevel SEM Modeling Software

With the advance of software packages, researchers now are more comfortable to build up multilevel models in their research practice (Hox, 2010; for comprehensive review of available software and packages, please refer to Goldstein, 2010; Snijders and Bosker, 2011). For example, an newly developed R package, xxM (Mehta, 2013), can be used to estimate multilevel SEM models featuring complex level-dependent data structures. The xxM is based on OpenMx (Boker et al., 2017) and a framework called n-Level Structural Equation Modeling (NL-SEM, e.g., Ryu and Mehta, 2017) which allows specifying multilevel models with observed and latent variables. Mplus (Muthén and Muthén, 1998-2017) and LISREL (Jöreskog et al., 2001) are commonly used structural equation modeling software for MCFAs. Researchers can use those software to examine the level-varying factorial structures, and simultaneously test different-level hypotheses (Muthén, 1994) with distinct model specifications. These programs present the overall model fit test statistics and fit indices with the provided multilevel SEM routines (e.g., TYPE = TWOLEVEL in Mplus), which cannot reveal possible misfit in respective levels. Instead, researchers can use partially saturated model (e.g., MAX MCFA in this study) or adjust the multi-group comparison approach (Muthén, 1994; Yuan and Bentler, 2007) to obtain level-specific model fit indices and test statistics in any SEM software. However, the programming of multilevel modeling practice would be intimidating to some researchers. Moreover, researchers can only specify the MCFA model with the same between- and within-level structure with Lisrel v.8 and below using SIMPLIS syntax via multi-group comparison (Jöreskog and Sörbom, 2004). To perform a MCFA with different factorial structures in the between and within level, researchers had to apply the LISREL syntax with matrix specification. Although the SURVEYGLIM procedure can be used to obtain the between-level and within-level covariance matrices after LISREL v.8.3 (Jöreskog et al., 2001), constructing MCFA using LISREL is still a daunting task which requires statistical computing operation in a multi-group comparison setting and LISREL coding in a matrix form.

Therefore, with the above-mentioned issues, researchers are in need of an effective and flexible multilevel modeling software which allows result comparison among competing models for optimal model selection.

Modeling Multilevel CFA Models Using iMCFA

Methodologists have provided suggestions and guidelines for constructing multilevel SEM models. Muthén (1994) proposed a stepwise procedure for multilevel model construction. He suggested that, in lack of model fit test and indices result for conventional CFA model, researchers should compute the intraclass correlation (ICC) measures for complex survey data. If the ICC value is nonzero or larger than certain thresholds (Muthén, 1994; Hedges and Hedberg, 2007; Hox, 2010), researchers should then build the multilevel model with respective within- and between-level structures. Hox (2002) suggested to compare the overall model-fit χ^2 test statistics of the one-level CFA (i.e., the null model) and of the independent MCFA (i.e., the MCFA with only variance estimates of between-level indicators and a hypothesized within-level model) with the saturated MCFA (i.e., MAX MCFA) as the first step to decide whether researchers should move on to establish a MCFA. Still other researchers (Yuan and Bentler, 2007; Ryu and West, 2009) provided level-specific model fit test statistics and fit indices to detect possible between level

[2] Balanced-design samples refer to equal sample size with respect to each group, whereas unbalanced-design samples refer to varying sample sizes in groups.

variation and potential model miss-specifications at respective levels.

The above-mentioned studies involved the design of the balanced synthetic dataset under the segregating approach (Yuan and Bentler, 2007) or the partially saturated model approach (Ryu and West, 2009) to capitalize on the importance of building adequate models with respective to the different level structures in analyzing the complex survey data. Constructing multilevel modeling according to the complex sampling design of the survey data could prevent erroneous inferences on the parameter estimates (Muthén, 1994; Yuan and Bentler, 2007), especially under a scenario with level-varying structures, in which the factor structure at the between level is different from that at the within level (Wu and Kwok, 2012). Besides, the precision of the inference of the relationship between items and factors, the scale reliability of constructs, and the variance explained of items would also be secured if we specify adequate multilevel models on the survey data (e.g., Raykov, 2004; Raykov and Marcoulides, 2006; Geldhof et al., 2013). Thus, we provide iMCFA using Lisrel v8.8 or below to help researchers build up valid multilevel models and to obtain ICC, variance-covariance matrix for separate level, and tabulated model comparison results on different modeling strategies[3].

In this study, we provide a general three-step procedure to construct a valid MCFA for complex survey data with level-varying structures and included the comparison of fit statistics and indices among three competing models for more precise model construction using iMCFA. At the first step, researchers should evaluate the model fit information of CFA as well as the congruency of the within-level parameter estimates between CFA and MAX MCFA. If CFA demonstrates bad overall model fit information or produces incongruent parameter estimates (especially the random effect estimates, e.g., factor variance or indicator residual variance) to the within-level of MAX MCFA, this is a strong message of potential between-level variation and level-varying structures. Next, researchers should focus on specifying within-level model using MAX MCFA by referencing to the model-fitχ^2 statistic and fit indices. After the within-level model is set, researchers can then proceed to construct a valid between-level model using MCFA based on the between-level model fit information at the third step with theoretical and empirical supports. An integrated MCFA program (iMCFA) is provided to manage the dataset and to perform the one-level CFA, MAX MCFA, and MCFA with Lisrel v.8 and below, which can aid the process of model selection and comparison. Two unbalanced and balanced datasets with level-varying structures (Study 1: the empirical unbalanced family IQ dataset with simple between and complex within structure; Study 2: the simulated balanced dataset with complex between and simple within structure) were used to illustrate the effectiveness and efficiency of the proposed approach and tool in building a valid MCFA model.

METHODS

We developed iMCFA to perform CFAs for complex survey data. The program is written in c++/CLI (Common language interface) in Microsoft Visual Studio 2012 (Microsoft Co. Ltd.). Researchers can use iMCFA to set the between- and within-level CFA models according to the theories and experiences, and to perform MCFA, CFA and/or MAX MCFA. Besides the computation of the level-varying variance-covariance matrices and automatic generation and execution of LISREL syntax, iMCFA will tabulate the results among the three competing models for further statistical decision making.

Performing CFA Analyses Using iMCFA

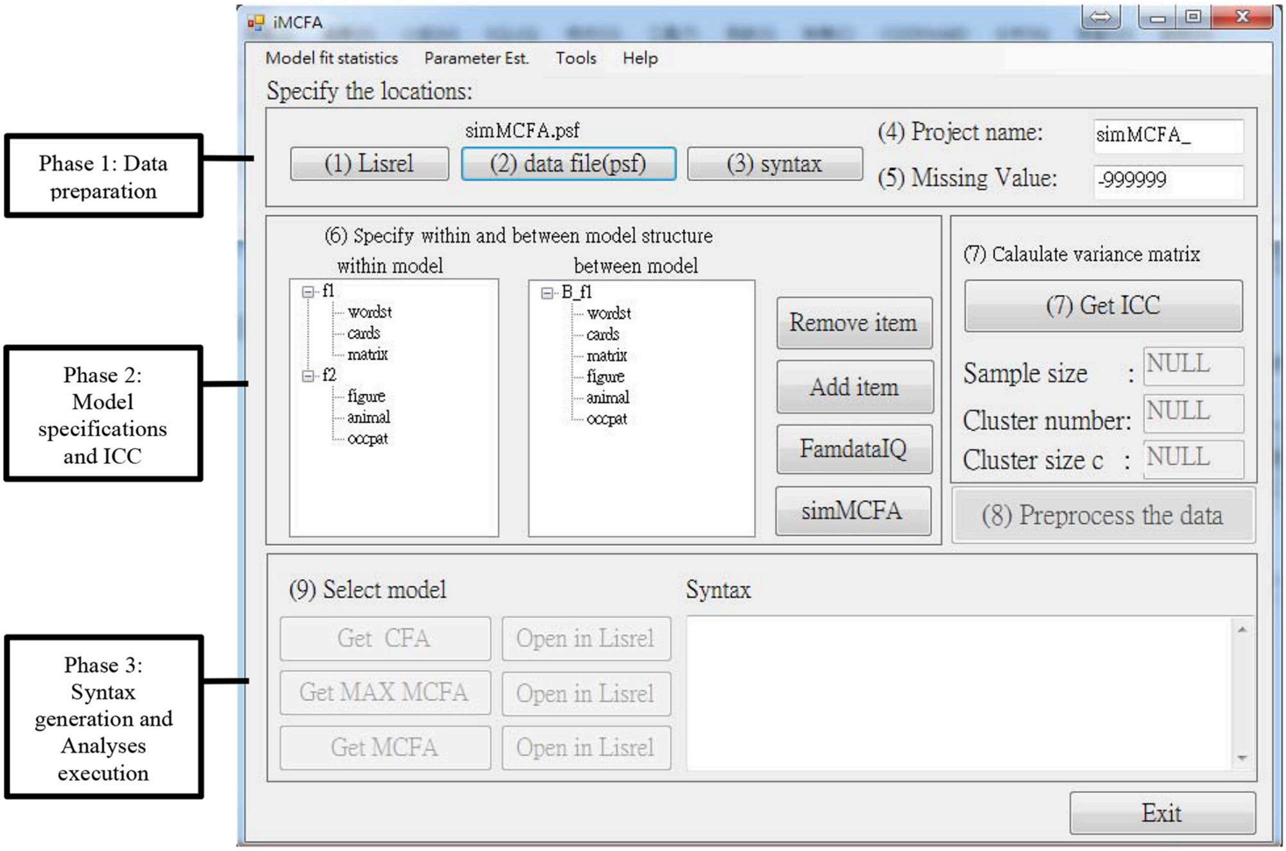
The snapshot of user interface of iMCFA is shown in **Figure 1**.

The interface of iMCFA is divided into three main phases, Phase 1: Data preparation, Phase 2: Model specifications and ICC calculation, and Phase 3: Syntax generation and Analyses execution. In the first phase, users need to specify the folder of LISREL program in Step (1), which contains LISREL*.exe, LisWin32.exe, and multilev*.exe, the folder of dataset file in Step (2) and the folder to save the generated syntax in Step (3). The default input data format for iMCFA is the LISREL data format (*.psf). Besides study variables, the imported dataset must include a Cluster ID variable and a Case ID variable, which identify the between-level and within-level sample units. The dataset should be sorted ascendingly according to Cluster ID and Case ID. iMCFA provides a tool to generate the Case ID variable. This tool can also convert the dataset from pure text format (*.dat) into LISREL data format (*.psf). The project name in Step (4) will be used as the prefix in the file name of all generated files, including the files of syntax, output, and variance-covariance matrix of variables. In Step (5), users have to specify a value of missing data (the default value will be−999999) to complete the data preparation phase.

In Phase 2 of model specification, Step (6) required researchers to specify the between- and within-level CFA structures. Users can add or remove latent and manifest variables after the data are read in. The variable labels can be edited and should be less than seven characters. After completing the above steps, click the 'Get ICC' button at Step (7) to do the ICC analysis and save information of sample size, cluster number, cluster size c, and variance-covariance matrix $\mathbf{S_{PW}}$, $\mathbf{S_B}$, $\mathbf{S_{B,MUML}}$ and $\mathbf{S_T}$ for the following analyses. At Step (8), the iMCFA gathers and saves all the matrices and values into the database for the following MAX MCFA, CFA, and MCFA analyses. For experienced researchers, these matrices can be used to conduct the multilevel analysis using any analytical programs with the distinct level-varying covariance matrices.

At Step (9) of Analysis and Syntax Phase 3, researchers can execute MAX MCFA, CFA, and MCFA separately. For MAX MCFA models, the within-level structure is based on the specification of the within-level model at Step (6), and the between-level structure is saturated, meaning all the between-level indicators are inter-correlated. For CFA models, the one-level structure is based on the specification of within-level model at Step (6). Researchers can specify MCFA model with unequal

[3] Users could utilize the generated V-C matrices and the modeling syntax from iMCFA in LISREL v. 9 to request for the analytical result and the diagram. However, due to major changes on the shell commands of LISREL v.9, the current procedure cannot be directly applied on LISREL 9 and its later versions.



(4) Project name: simMCFA_

(5) Missing Value: -999999

Sample size : NULL

Cluster number: NULL

Cluster size c : NULL

FIGURE 1 | Snapshot of iMCFA interface with an example of two-level CFA model with one factors at between level and two factors at within level.

between- and within-level structures using iMCFA. For example, we specify a MCFA model (shown in **Figure 2**) with one between-level factor and two within-level factors. To correctly perform the MCFA, researchers need to keep the order and the number of items the same in between- and within-level model specified at step (6).

LISREL syntax for three models will be generated and executed after clicking the buttons of "Get MAX MCFA," "Get CFA," and "Get MCFA." The model goodness-of-fit test statistics and indices and parameter estimates will be retrieved from the output at this step. The syntax will be presented in the bottom-right text box. For convenience, users can click the "Open in LISREL" button to execute the corresponding syntax in LISREL, which can generate the analytic result and the model diagram.

On the top of the panel, researchers can request the tabulated analysis results by clicking tabs of "Model fit statistics" and "Parameter Est." The Model fit statistics tab shows the fit test statistic and fit index information for three models. The Parameter Est. tab summarizes the estimates of factor loadings, residual variances of each item, and the covariances among latent factors for three models. Researchers are allowed to save these tables in a text file, which can be found in the same folder of syntax files.

In the following sections, we used two datasets to demonstrate iMCFA, one was a family IQ dataset (famdataIQ.psf, Hox, 1993, 2010) and the other was a simulated dataset (simMCFA.psf, Wu and Kwok, 2012). The commonly-used criteria of model fit indices were used to assess the goodness of fit of the proposed models to the dataset.

Study 1: Empirical Unbalanced Dataset With Simple Between and Complex Within Structure

Data Description

The empirical dataset (famdataIQ.psf) is from the dissertation study of Van Peet (1992) which also appeared in Hox (1993, 2010). Data were collected on 400 children nested within 60 families with a minimum 4 and maximum 12 children in each family ($M = 6.67$). The instrument used was the Groninger Intelligent Test (GIT) which consisted of six subtests, including wordlist, laying cards, matrices, hidden figures, naming animals, and naming occupations. Strong correlation among members in a family were expected because intelligence is assumed to be greatly influenced by heredity and environments. Scores on the six subtests for this hierarchical data were then divided into family level and individual level variables. According to Hox (1993), there was a common factor in the family level

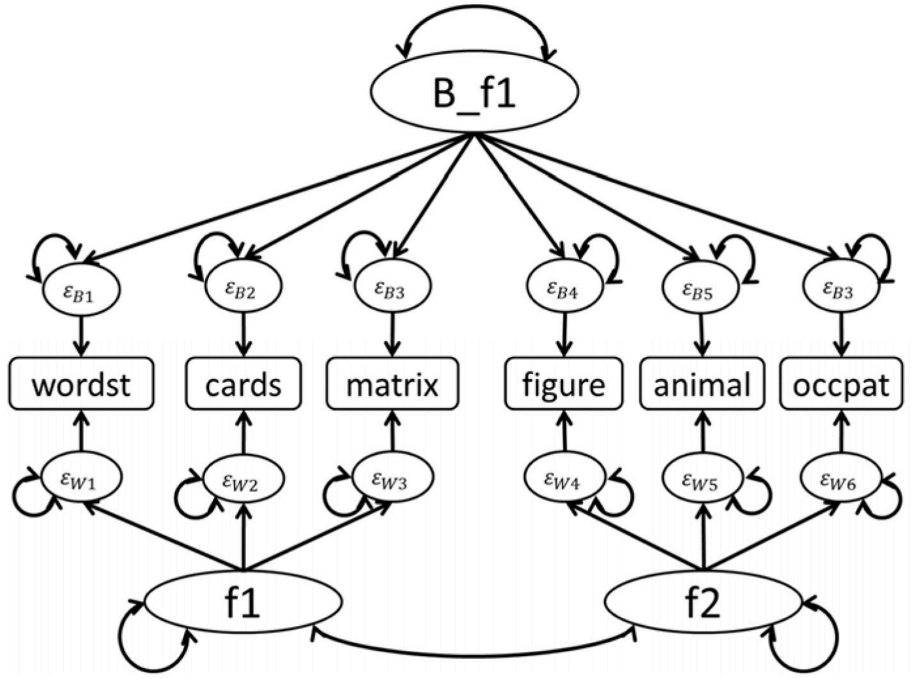


FIGURE 2 | MCFA model with 1 factor at between level and 2 factors at within level for FamdataIQ dataset.

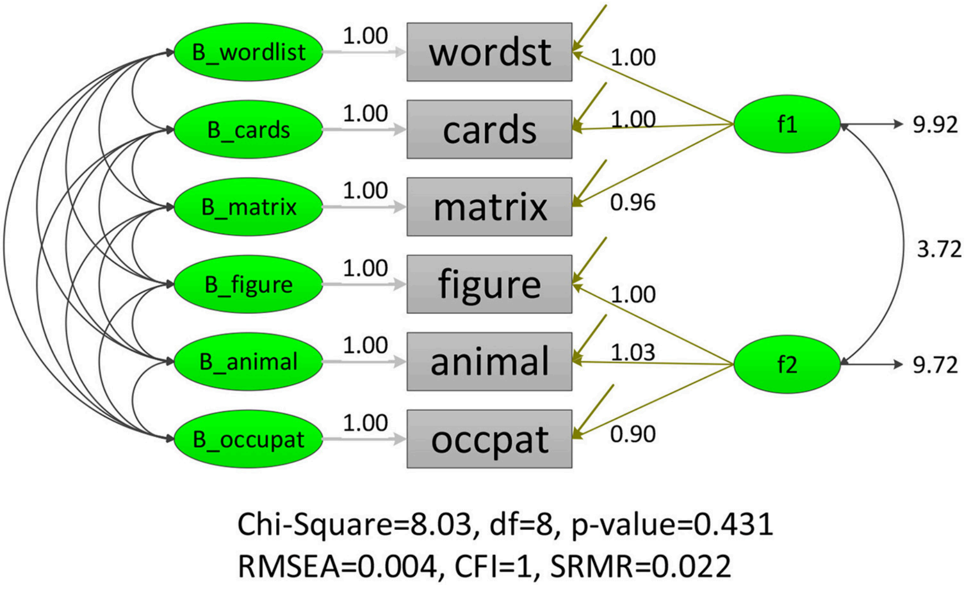


FIGURE 3 | LISREL illustration of MAX MCFA model on FamdataIQ dataset.

due to shared genetic and environmental influences while in the individual level two separate factors existed to explain the idiosyncrasy in each individual's intelligence.

Model Specification

The famdataIQ.psf involved eight variables: family id, user id, wordlist (wordst), laying cards (cards), matrices (matrix), hidden figures (figure), naming animals (animal), and naming occupations (occpat). We constructed two factors (f1 and f2) in the within level and one factor (B_f1) in the between level based on Hox (1993). In the within level, wordst, cards, and matrix were loaded on f1, while figure, animal, and occpat were loaded on f2. In the between level, all six items were loaded on B_f1 as shown in **Figure 2**.

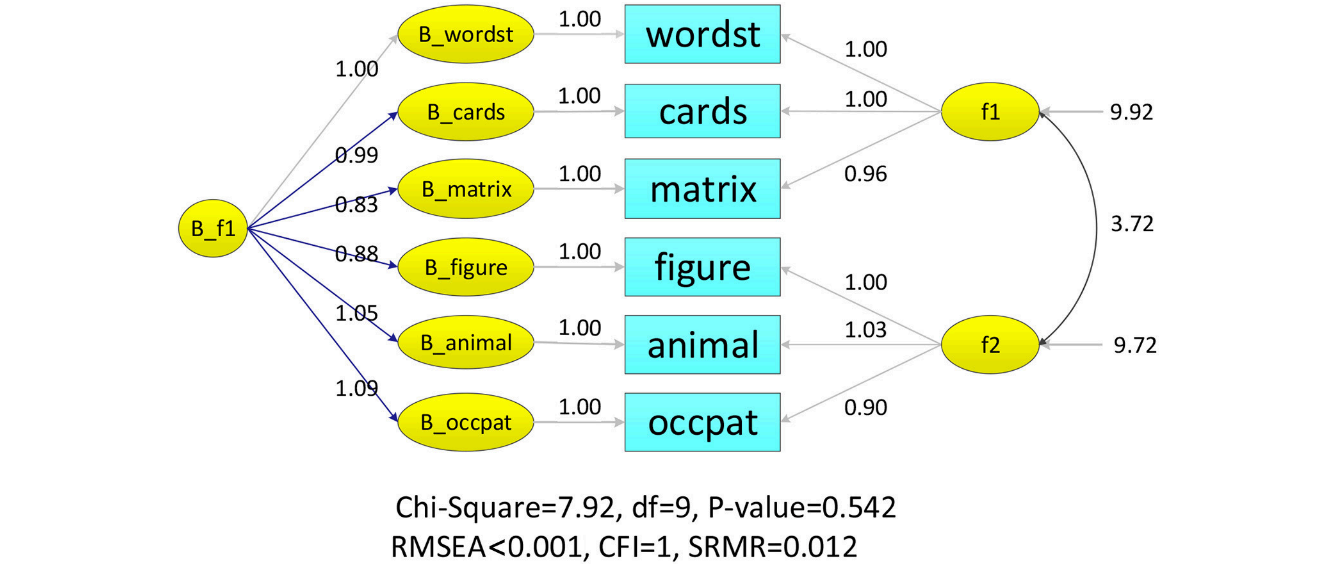


FIGURE 4 | LISREL illustration of MCFA model on famdataIQ dataset.

Result for Study 1

Three modeling techniques in iMCFA were applied to analyze the family IQ dataset. All models had adequate model-fit test statistic and fit indices for the data with unequal family- and individual-level structure (e.g., for MCFA, $\chi^2 = 7.920$ with $df = 9$, CFI = 1.000, RMSEA < 0.001, SRMR = 0.012; for MAX MCFA, $\chi^2 = 8.027$ with $df = 8$, CFI = 1.000, RMSEA = 0.004, SRMR = 0.022; for CFA, χ^2 =10.241 with $df = 8$, CFI = 0.999, RMSEA = 0.027, SRMR = 0.016;). The ICCs for six indicators were larger than 0.369 (as shown in **Table 1**), which indicated potential between-level variation and the need to use multilevel CFA techniques (Hox, 2010). The path diagrams with analytical result of MCFA and MAX CFA are illustrated in **Figure 3**, **4**. The results of MCFA confirmed the existence of a general between-level intelligence construct, which could explain the influence of heredity and environment in a family. The three modeling techniques exhibited a 2-factor structure in the within-level model representing the idiosyncrasies in each individual's intelligence (Van Peet, 1992; Zimprich and Martin, 2009). We then compared the performance of these three modeling techniques on this complex survey dataset.

The analysis result of three modeling techniques was tabulated in **Table 2**. The MAX MCFA model yielded similar model evaluation result and parameter estimates as the MCFA in the within level. However, when CFA was applied on this family dataset, the factor loading estimates were statistically different from those of the MAX MCFA model or MCFA [e.g., $\hat{\lambda}^{MCFA}_{occupats,W\_IQ2} = \hat{\lambda}^{MAX}_{occupats,W\_IQ2} = 0.901$, vs. $\hat{\lambda}^{CFA}_{occupats,W\_IQ2} = 1.071$, $t_{(df)} = 1.99(798)$, $p = 0.046$]. The relative difference of factor loading estimates for CFA compared to the MAX MCFA and MCFA ranged from −6.18 to 15.87%, which could be considered as a moderate to substantial difference (Flora and Curran, 2004). This level of incongruence between parameter estimates of CFA and MAX MCFA might indicate the necessity of constructing a multilevel model with level-varying structures for this dataset.

TABLE 1 | ICC and R^2-values of indicators of three models for FamdataIQ dataset ($N = 400$, $G = 60$).

	Wordst	Cards	Matrix	Figure	Animal	Occpat
ICC	0.399	0.408	0.369	0.374	0.419	0.503
R^2_MCFA						
Individual-level	0.614	0.651	0.589	0.588	0.678	0.596
Family-level	0.885	0.874	0.781	0.787	0.937	0.880
R^2_ CFA	0.735	0.733	0.657	0.649	0.789	0.737
R^2_MAX CFA	0.614	0.651	0.589	0.588	0.678	0.596

As for the random effect estimate, CFA generated an overall estimate of factor variance, which was roughly the summation of the family- and individual-level variance components (e.g., $\Psi^{CFA}_{wordst} = 7.131$ equals $\Psi^{MCFA}_{wordst,W} = 6.228$ plus $\Psi^{MCFA}_{wordst,B} = 1.024$) and the MAX MCFA yielded consistent individual-level factor variance to the MCFA (e.g., $\Psi^{MCFA}_{wordst,W} = \Psi^{MAX}_{wordst,W} = 6.228$). However, CFA tended to generate inflated R^2 for the within-level indicators compared to MCFA and MAX MCFA.

We also compared the parameter estimates of three proposed CFA modeling techniques using iMCFA with MUML variance decomposition and those obtained from Mplus 6.11 with FIML estimation (as shown in Table S.1 in the Appendix). For the parameter estimates in the within-level model, the averaged relative bias was 0.022% (SD = 0.543%) for MCFA, 0.040% (SD = 0.126%) for MAX MCFA, and 1.037% (SD = 1.827%) for CFA. The relative bias of estimates between these two programs for multilevel CFAs could be deemed as trivial (Flora and Curran, 2004).

TABLE 2 | Three CFA models of empirical famdataIQ dataset. ($N = 400$, $G = 60$).

	MCFA		CFA		MAX MCFA	
Chi-square (df)	7.920(9)		10.241(8)		8.027(8)	
CFI	1.000		0.999		1.000	
RMSEA	0.000		0.027		0.004	
SRMR	0.012		0.016		0.022	
	Est.	**SE**	**Est.**	**SE**	**Est.**	**SE**
INDIVIDUAL LEVEL						
W_IQ1 by						
wordlist	1		1		1	
cards	1.001***	0.069	0.979***	0.049	1.001***	0.069
matrices	0.962***	0.068	0.906***	0.048	0.962***	0.068
W_IQ2 by						
figures	1		1	0	1	0
animals	1.026***	0.071	1.093***	0.056	1.026***	0.071
occupats	0.901***	0.064	1.071***	0.056	0.901***	0.064
Cov(W_IQ1,W_IQ2)	3.721***	0.658	12.622***	1.344	3.721***	0.658
Var(W_IQ1)	9.918***	1.173	19.755***	1.937	9.918***	1.173
Var(W_IQ2)	9.724***	1.179	17.136***	1.828	9.724***	1.179
RESIDUAL VAR						
wordlist	6.228***	0.677	7.131***	0.799	6.228***	0.677
cards	5.335***	0.637	6.881***	0.768	5.355***	0.637
matrices	6.414***	0.659	8.483***	0.800	6.414***	0.659
figures	6.824***	0.696	9.286***	0.840	6.824***	0.696
animals	4.859***	0.625	5.489***	0.703	4.859***	0.625
occupats	5.358***	0.556	7.016***	0.757	5.358***	0.556
FAMILY LEVEL						
B_IQ by						
wordlist	1					
cards	0.985***	0.083				
matrices	0.831***	0.080				
figures	0.878***	0.109				
animals	1.050***	0.113				
occupats	1.091***	0.118				
Var(B_IQ)	9.677***	1.797				
RESIDUAL VAR						
wordlist	1.024ns	0.760				
cards	1.449*	0.728				
matrices	1.947*	0.767				
figures	2.161**	0.813				
animals	0.495 ns	0.662				
occupats	1.763*	0.759				

$p < 0.05^{*}$, $p < 0.01^{**}$, $p < 0.001^{***}$.

χ^2, Chi-square value; df, Degrees of freedom; CFI, Comparative fit index; RMSEA, Root mean square error of approximation; SRMR, Standardized root mean square residual. The normal font denotes the fixed effect and intercept estimate; the italic denotes the random effect estimate.

Study 2: Simulated Balanced Dataset With Complex Between and Simple Within Structure

Data Description

The simMCFA.psf involved nine indicators (V1 to V9). In the population model, all nine indicators were loaded on one factor (W_f1) at the within level and three factors (B_f1, B_f2, and B_f3) at the between level. This simulated balanced dataset was generated using Monte Carlo procedure of Mplus 6.11 with 10,000 observations nested within 50 groups (i.e., each group had 200 participants). All factor loadings were set at 0.80, and the residual variances of outcome variables were fixed at 0.36 in both within- and between-level models. Moreover, covariances among three between-level latent factors were fixed at 0.30. The ICCs in **Table 3** for nine indicators were larger than 0.388. The detailed settings of the true model with cross-loaded factor loadings could be referred to scenario 3 in Wu and Kwok (2012).

Result for Study 2

The same three modeling techniques with a simple structure in the within level were applied for the simulated dataset. The analysis results were tabulated in **Table 4**. Furthermore, we also constructed a misspecified MCFA with one factor in both between- and within-level model, that is, the between-level model did not confirm to the true three-factor structure. Likewise, the correctly-specified MCFA and MAX MCFA models yielded similar model evaluation results and parameter estimates; however, CFA yielded inadequate overall model-fit test statistic and fit indices (For CFA, overall $\chi^2 = 12699.87$ with $df = 27$, CFI = 0.881, RMSEA = 0.217 SRMR = 0.090; For MAX MCFA, within-level $\chi^2 = 26.089$ with $df = 27$, CFI = 1.000, RMSEA < 0.001, SRMR = 0.003; For MCFA, the between-level $\chi^2 = 824.5$ with $df = 24$, CFI = 0.991, RMSEA = 0.058, SRMR = 0.027). When the CFA was applied, the factor loading estimates deviated from those of the MCFA and the MAX MCFA [e.g., $\hat{\lambda}^{MCFA}_{V5,f1} = 0.802$ and $\hat{\lambda}^{MAX}_{V5,f1} = 0.802$ vs. $\hat{\lambda}^{CFA}_{V2,f1} = 0.603$, $t_{(df)} = 15.63(19998)$, $p < 0.001$], and the relative bias of factor loading estimates for the CFA comparing to the MCFA and MAX MCFA ranged from −38.47 to 1.20%, which could be seen as trivial to substantial differences (Flora and Curran, 2004). Both the model lack-of-fit information and the incongruence of parameter estimates inform the need of further multilevel modeling with level-varying structures. Researchers can use maximum modeling techniques with the within-level model goodness-of-fit tests and indices to construct a valid within-level model, and proceed to use MCFA with the respective between-level model-fit information to have a valid between-level model.

Comparing the correctly- and miss-specified MCFAs, the model-fit χ^2 statistic indicated that the misspecified MCFA model did not fit the data exactly, and the fit indices exhibited more severe model badness-of-fit result (For misspecified MCFA on the left hand side of **Table 4**, CFI = 0.931, RMSEA = 0.171, and SRMR = 0.204). With the valid within-level structure at the MAX MCFA step, researchers can then build up several competing MCFAs with different between-level models and conduct the model comparison analyses with the aid of model-fit χ^2 and fit indices provided in the MCFA step of iMCFA to select the proper multilevel model with statistical and theoretical evidence.

CFA tended to generate smaller R^2 for the within-level indicators compared to the MCFA and MAX MCFA models as shown in **Table 3**. We compared the parameter estimates of iMCFA with MUML and those of Mplus 6.11 with FIML (as shown in Table S.2 in the Appendix). For the individual-level model, the relative bias of MCFA, MAX MCFA, and CFA was

TABLE 3 | ICC and R^2-values of nine indicators of three models for simMCFA dataset.

	V1	V2	V3	V4	V5	V6	V7	V8	V9
ICC	0.516	0.516	0.535	0.487	0.388	0.484	0.353	0.470	0.456
R^2_MCFA									
Within-level	0.642	0.641	0.636	0.637	0.640	0.638	0.633	0.635	0.639
Between-level	0.737	0.675	0.783	0.639	0.675	0.420	0.662	0.593	0.818
R^2_ CFA	0.616	0.577	0.601	0.468	0.439	0.372	0.424	0.421	0.434
R^2_ MAX_MCFA	0.642	0.641	0.636	0.637	0.640	0.638	0.633	0.635	0.639

very close to zero. Indicating the parameters estimates generated by the MUML were consistent with those generated by the FIML estimator.

DISCUSSION AND CONCLUSION

In order to reduce the complexity of using multilevel CFA techniques, we provided the iMCFA program as an integrated tool to manage three most commonly-used CFA modeling techniques, namely regular CFA, MAX MCFA, and MCFA, on a user-friendly interface to analyze complex survey data with LISREL v.8 and below. The capacity to specify level-varying structures is fundamental to ensure the accuracy of analytical results in various CFA analyses with complex survey dataset. Failing to build up a model conforming to the true multilevel data structure may lead to erroneous analytical results and incorrect conclusions (Wu and Kwok, 2012). Even with the advance of the analytical software on analyzing various SEM models, it is difficult for researchers to specify MCFA models with level-varying structures with the supports of model-fit test and fit indices. Specifically, there is still not an efficient function in these programs to compare the adequacy of different modeling techniques simultaneously on the multilevel data. In this study, we used the iMCFA program to compare the performance of MCFA, miss-specified MCFA, MAX MCFA and CFA on two different datasets (one empirical unbalanced and one simulated balanced dataset) considering their level-varying structure and balanced/unbalanced design.

The different analytical results of CFA compared with the MAX MCFA technique may indicate a potential between-level structure in the dataset. In the illustrations, we demonstrated that when the relative bias of within-level factor loading estimates of CFA and MAX MCFA was moderate to substantial (Flora and Curran, 2004), there could be level-varying structures in the complex survey data. For the multilevel dataset with level-varying structures, CFA generated conflated parameter estimates of fixed and random effects with overall variance-covariance matrix along with the inconsistent standard error estimates. Besides, due to the conflated estimates in the one-level modeling, the variance explained measures (e.g., R^2) of CFA were different from the outputs of MCFA and MAX MCFA (Wu and Kwok, 2012; Geldhof et al., 2013; Wu et al., 2017). For a complex survey dataset, the association between the R^2 generated by regular CFA and the R^2 measures in respective between- and within-levels by MCFA models warrants future simulation and/or mathematical investigations.

With level-specific variance components, MCFA could only generate consistent results when the analytical model is close to the true multilevel structures in both between- and within-level models simultaneously; while with a saturated between-level model, MAX MCFA model could be utilized to construct the individual model consistent with the within-level structure of the true multilevel model. If researchers and practitioners fail to use modeling techniques that are congruent with the multilevel structure of the complex survey data, they should exert caution in interpreting or making inferences from a regular CFA and MCFA. Instead, researchers would benefit from the use of MAX MCFA model offered in iMCFA in dealing with the complex survey data. If researchers are interested in only the research question about the within-level model, they should use the result from MAX MCFA analysis to draw a conclusion for the variation of within-level sampling units.

If researchers aim to answer research questions concerning different levels of the dataset, they could start with a MAX MCFA model to build an optimal within-level model. Next, they could go further to specify their between-level structure using MCFA to capture the between-level variation in their complex survey dataset (Hox, 2002). The model-fit information are indicators of the quality of hypothesized between-level model.

To complete the above-mentioned steps for building up an adequate multilevel CFA model, researchers or practitioners can use iMCFA to conduct multilevel CFA with equal or unequal between- and within-level structures in an effective and efficient way. They can use the tabulated analytical results provided by iMCFA to compare the performance of the three modeling techniques and to select the optimal model for statistical inference. Researchers can further use the generated LISREL syntax to request model diagrams and perform more detailed and advanced analyses in LISREL v.8 and below. The generated LISREL syntax of the MAX MCFA model and the MCFA for familyIQ dataset is provided in the Appendix. We also performed the equality check for the analytical result of iMCFA with the proposed algorithm. The within-level fixed-effect parameter estimates of target model generated by iMCFA were consistent with Mplus[4], which is one of the most commonly used SEM software.

[4] In study 2, the simulation dataset was generated in Mplus with FIML estimation. The analytical result were congruent in the within-level fixed-effect estimates between two programs. The noticeable differences between some of the between-level fixed- and random-effects (in Table S.2) would result from the different estimation methods. More simulation studies with comprehensive experimental designs should be conducted to thoroughly investigate the performance of different modeling techniques, estimation methods, and statistical programs on analyzing complex survey data.

TABLE 4 | Fit information and parameter estimates of hypothesized and misspecified models on dataset.

		MCFA		**MISS MCFA**		**CFA**		**MAX MCFA**	
Chi-square (df)		824.499(24)		7897.358(27)		12699.87(27)		26.089(27)	
CFI		0.991		0.931		0.881		1.000	
RMSEA		0.058		0.171		0.217		0.000	
SRMR		0.027		0.204		0.090		0.003	
		Est.	**SE**	**Est.**	**SE**	**Est.**	**SE**	**Est.**	**SE**
WITHIN LEVEL									
W_f1 by	V1	0.800	—	0.800	—	0.800	—	0.800	—
	V2	0.799	0.009	0.799	0.009	0.774	0.010	0.799	0.009
	V3	0.792	0.009	0.792	0.009	0.801	0.010	0.792	0.009
	V4	0.791	0.009	0.791	0.009	0.673	0.010	0.791	0.009
	V5	0.802	0.009	0.802	0.009	0.603	0.009	0.802	0.009
	V6	0.794	0.009	0.794	0.009	0.599	0.010	0.794	0.009
	V7	0.798	0.009	0.798	0.009	0.576	0.009	0.798	0.009
	V8	0.787	0.009	0.787	0.009	0.625	0.009	0.787	0.009
	V9	0.798	0.009	0.798	0.009	0.633	0.009	0.798	0.009
Var(W_f1)		1.001	0.021	1.001	0.021	1.985	0.044	1.001	0.021
RESIDUAL VAR									
	V1	0.357	0.006	0.357	0.006	0.791	0.014	0.357	0.006
	V2	0.358	0.006	0.358	0.006	0.872	0.015	0.358	0.006
	V3	0.359	0.006	0.359	0.006	0.848	0.014	0.359	0.006
	V4	0.358	0.006	0.358	0.006	1.020	0.016	0.358	0.006
	V5	0.362	0.006	0.362	0.006	0.920	0.014	0.362	0.006
	V6	0.358	0.006	0.358	0.006	1.205	0.018	0.358	0.006
	V7	0.369	0.006	0.369	0.006	0.895	0.014	0.369	0.006
	V8	0.356	0.006	0.356	0.006	1.066	0.016	0.356	0.006
	V9	0.360	0.006	0.360	0.006	1.037	0.016	0.360	0.006
BETWEEN LEVEL									
B_f1 by	V1	0.800	—	0.800	—				
	V2	0.802	0.015	1.191	0.099				
	V3	0.879	0.017	1.075	0.092				
B_f2 by	V4	0.800	—	2.196	0.174				
	V5	0.601	0.019	1.557	0.122				
	V6	0.566	0.019	1.592	0.128				
B_f3 by	V7	0.800	—	1.894	0.199				
	V8	0.973	0.025	2.198	0.224				
	V9	1.139	0.031	2.496	0.250				
Var(B_f1)		1.253	0.045	0.051	0.009				
Cov(B_f1,B_f2)		0.552	0.031						
Cov(B_f1,B_f3)		1.025	0.048						
Var(B_f2)		0.305	0.022						
Cov(B_f2,B_f3)		0.036	0.021						
Var(B_f3)		0.538	0.029						
Residual Var									
	V1	0.292	0.014	0.593	0.015				
	V2	0.347	0.014	0.620	0.016				
	V3	0.240	0.015	0.634	0.016				
	V4	0.270	0.018	0.400	0.015				
	V5	0.238	0.012	0.379	0.013				

(Continued)

TABLE 4 | Continued

	MCFA		MISS MCFA		CFA	MAX MCFA
V6	0.569	0.016	0.672	0.017		
V7	0.193	0.011	0.195	0.011		
V8	0.343	0.014	0.338	0.014		
V9	0.149	0.016	0.199	0.014		

(N of sample = 10,000 with Group Number = 50 and Group Size = 200).

All above parameter estimates are statistically significant at the level of $p < 0.05$.

χ^2, Chi-square value; df, Degrees of freedom; CFI, Comparative fit index; RMSEA, Root mean square error of approximation; SRMR, Standardized root mean square residual. The normal font denotes the fixed effect and intercept estimate; the italic denotes the random effect estimate. MCFA: the multilevel CFA model with one within-level factor and three between-level factors as the population model of simulation dataset. MISS MCFA: the miss-specified multilevel CFA model with one within-level factor and one between-level factor. CFA: the miss-specified CFA model with a uni-factor single-level structure. MAX MCFA: the multilevel CFA model with one within-level factor and saturated between-level structure.

In sum, when analyzing complex survey data with level-varying structures, we recommend researchers in the applied areas to use iMCFA to simultaneously perform their analyses with the three proposed modeling techniques. The MAX MCFA model answers research questions about the within-level sampling units, and serves as the baseline for further MCFA construction in response to the level-varying questions for both levels. The factor scores of multilevel measurement analysis from iMCFA could be incorporated in the structure model as the 2-step approach (Anderson and Gerbing, 1988, 1992) to conduct the multilevel SEM analysis. Our illustrations demonstrated that iMCFA can help researchers in their empirical and theoretical study to perform multilevel analyses on complex survey data.

System Requirement, Functionality, and Future Development of iMCFA

iMCFA requires 15 MB of hard disk space to store and has been developed and tested on Windows 7/8/10 32 bits and 64 bits operation systems with LISREL version v8.7 or v8.8 installed. Executing time will vary depending on the complexity of users' model. To consider the computation loading, for the current version of iMCFA, we set the limit of the maximal number of factors as 10, and the maximal number of items as 100.

The iMCFA tool focuses on integrating the functionalities with respect to performing multilevel confirmatory factor analysis with simple or complex structures. By default, iMCFA sets the first indicator of each factor to be the marker variable (e.g., the wordst for f1 and the figure for f2 in **Figure 1**). Users could re-arrange the input sequence of variables to set the markers. The current version allows only the indicators of continuous scale. Users could set up missing flag in Phase 1 to mark the missingness. To utilize the MUML estimation with multi-group comparison analysis in Lisrel, iMCFA uses the pair-wise deletion for incomplete data to compose level-specific variance-covariance matrices for the complex survey data. With the assumption of Missing at Random (MAR), users can process the incomplete dataset with multiple imputation procedure (Little and Rubin, 1987; Enders, 2010) prior to the use of iMCFA. Users can also revise the generated syntax of three modeling techniques from iMCFA to utilize Full Information Maximum Likelihood method (FIML, Arbuckle, 1996), the default estimation method in Lisrel, for their incomplete raw data with missing values. The equality constraints or fitting multiple-group models are not allowed in the current version of iMCFA.

The function to specify the factor loadings of cross-loaded items and correlated item residuals, and the feature of parameter comparisons with Wald test (Wald, 1943) and family-wise Type-I error rate control among three models will be provided in the following version. In addition, standard errors, *t* values, and significance of corresponding parameters are not included in the iMCFA tabulated output because the output focuses on comparing the model fit of the three modeling techniques. Nonetheless, researchers can use the generated syntax to request the information from LISREL.

Though we applied iMCFA on various types of datasets and models in this study, for more general cases of balanced or unbalanced complex survey data with level-varying structures, the performance of different estimation methods, the sensitivity of level-specific model-fit test statistics and fit indices in detecting lack-of-fit in multilevel CFA and SEM analysis still need more investigation using simulation and empirical approaches.

AUTHOR CONTRIBUTIONS

J-YW designed the study, conducted the literature review and took a leading role in writing the manuscript. Y-HL helped with data analysis and literature review. JL helped with the programing.

FUNDING

This research was partially supported by Grant 104-2628-H-009-001-SS3 and 105-2511-S-009-009-MY3 from the Ministry of Science and Technology, Taiwan.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2018.00251/full#supplementary-material

REFERENCES

Anderson, J. C., and Gerbing, D. W. (1988). Structural equation modeling in practice: a review and recommended two-step approach. *Psychol. Bull.* 103, 411–423. doi: 10.1037/0033-2909.103.3.411

Anderson, J. C., and Gerbing, D. W. (1992). Assumptions and comparative strengths of the two-step approach: comment on Fornell and Yi. *Sociol. Methods Res.* 20, 321–333. doi: 10.1177/0049124192020003002

Arbuckle, J. L. (1996). "Full information estimation in the presence of incomplete data," in *Advanced Structural Equation Modeling: Issues and Techniques*, eds G. A. Marcoulides and R. E. Schumacker (Mahwah, NJ: Lawrence Erlbaum Associates), 243–277.

Boker, S. M., Neale, M. C., Maes, H. H., Wilde, M. J., Spiegel, M., Brick, T. R., et al. (2017). *OpenMx: Extended Structural Equation Modelling* (Version 2.8.3). Available online at: https://cran.r-project.org/web/packages/OpenMx/index.html

Bollen, K. A. (1989). *Structural Equation Models with Latent Variables*. Hoboken, NJ: Wiley Interscience.

Bovaird, J. A. (2007). "Multilevel structural equation models for contextual factors," in *Modeling Contextual Effects in Longitudinal Studies*, eds T. D. Little, J. A. Bovaird, and N. A. Card (Mahwah, NJ: Lawrence Erlbaum Associates), 149–182.

Enders, C. K. (2010). *Applied Missing Data Analysis, 1st Edn.* New York, NY: The Guilford Press.

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

Geldhof, G. J., Preacher, K. J., and Zyphur, M. J. (2013). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychol. Methods* 19, 72–91. doi: 10.1037/a0032138

Goldstein, H. (2010). *Multilevel Statistical Models*, 4th Edn. John Wiley & Sons, Ltd. Available online at: http://onlinelibrary.wiley.com/book/10.1002/9780470973394

Heck, R. H., and Thomas, S. L. (2008). *An Introduction to Multilevel Modeling Techniques, 2nd Edn.* New York, NY: Routledge.

Hedges, L., and Hedberg, E. C. (2007). Intraclass correlations for planning group randomized experiments in rural education. *J. Res. Rural Edu.* 22, 1–15. doi: 10.3102/0162373707299706

Hox, J. J. (1993). "Factor analysis of multilevel data: Gauging the Muthén model," in *Advances in Longitudinal and Multivariate Analysis in the Behavioral Sciences*, eds J. H. L. Oud and R. A. W. van Blokland-Vogelesang (Nijmegen: ITS), 141–156.

Hox, J. J. (2002). *Multilevel Analysis Techniques and Applications*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hox, J. J. (2010). *Multilevel Analysis: Techniques and Applications, 2nd Edn.* New York, NY: Routledge Academic.

Hox, J. J., and Maas, C. J. M. (2004). "Multilevel structural equation models: The limited information approach and the multivariate multilevel approach," in *Recent Developments on Structural Equation Models: Theory and Applications*, eds K. van Montfort, J. Oud, and A. Satorra (Kluwer Academic Publishers), 135–149.

Jöreskog, K. G., and Sörbom, D. (2004). *LISREL 8.7 for Windows [Computer Software]*. Lincolnwood, IL: Scientific Software International, Inc.

Jöreskog, K. G., Sörbom, D., and Du Toit, S. H. C. (2001). *LISREL 8: New Statistical Features*. Lincolnwood, IL: Scientific Software International.

Kaplan, D. W. (2008). *Structural Equation Modeling: Foundations and Extensions*. Thousand Oaks, CA: Sage Publications.

Kline, R. B. (2016). *Principles and Practice of Structural Equation Modeling, 4th Edn.* New York, NY: The Guilford Press.

Little, R. J. A., and Rubin, D. B. (1987). *Statistical Analysis with Missing Data, 2nd Edn*. Hoboken, NJ: Wiley-Interscience.

Mehta, P. D. (2013). *xxM User's Guide.* Available online at: http://xxm.times.uh.edu/

Mehta, P. D., and Neale, M. C. (2005). People are variables too: multilevel structural equations modeling. *Psychol. Methods* 10, 259–284. doi: 10.1037/1082-989X.10.3.259

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika* 54, 557–585. doi: 10.1007/BF02296397

Muthén, B. O. (1990). *Mean and covariance structure analysis of hierarchical data. Presented at the Psychometric Society*, Princeton, NJ. Available online at: http://escholarship.org/uc/item/1vp6w4sr

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *J. Edu. Meas.* 28, 338–354. doi: 10.1111/j.1745-3984.1991.tb00363.x

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociol. Methods Res.* 22, 376–398. doi: 10.1177/0049124194022003006

Muthén, B. O., and Satorra, A. (1989). "Multilevel aspects of varying parameters in structural models," in *Multilevel Analysis of Educational Data*, ed R. D. Bock (San Diego, CA: Academic Press), 87–99.

Muthén, B. O., and Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociol. Methodol.* 25, 267–316. doi: 10.2307/271070

Muthén, L. K., and Muthén, B. O. (1998-2017). *Mplus User's Guide. 7th Edn.*, Los Angeles, CA: Muthén & Muthén.

Raykov, T. (2004). Behavioral scale reliability and measurement invariance evaluation using latent variable modeling. *Behav. Ther.* 35, 299–331. doi: 10.1016/S0005-7894(04)80041-8

Raykov, T., and Marcoulides, G. A. (2006). On multilevel model reliability estimation from the perspective of structural equation modeling. *Struct. Equat. Model.* 13, 130–141. doi: 10.1207/s15328007sem1301_7

Ryu, E., and Mehta, P. (2017). Multilevel factorial invariance in n-level structural equation modeling (nSEM). *Struct. Equat. Model. A Multidiscipl. J.* 24, 936–959. doi: 10.1080/10705511.2017.1324311

Ryu, E., and West, S. G. (2009). Level-specific evaluation of model fit in multilevel structural equation modeling. *Struct. Equat. Model. A Multidiscipl. J.* 16, 583–601. doi: 10.1080/10705510903203466

Searle, S. R., Casella, G., and McCulloch, C. E. (1992). *Variance Components*, *1st Edn*. New York, NY: Wiley-Interscience.

Snijders, T. A. B., and Bosker, R. (2011). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*, *2nd Edn*. London: SAGE Publications Ltd.

Stapleton, L. M. (2006). "Using multilevel structural equation modeling techniques with complex sample data," in *Structural Equation Modeling: A Second Course*, eds G. R. Hancock and R. O. Mueller (Greenwich, CT: Information Age Publishing), 345–383.

Stapleton, L. M. (2008). "Analysis of data from complex surveys," in *International Handbook of Survey Methodology*, eds E. D. de Leeuw, J. J. Hox, and D. A. Dillman (New York, NY: Lawrence Erlbaum Associates), 342–369.

Thompson, B. (2004). *Exploratory and Confirmatory Factor Analysis: Understanding Concepts and Applications*. Washington, DC: American Psychological Association.

Van Peet, A. A. (1992). *De Potentieeltheorie van Intelligentie (The Potentiality Theory of Intelligence)*. PhD dissertation, University of Amsterdam.

Varin, C., and Vidoni, P. (2005). A note on composite likelihood inference and model selection. *Biometrika* 92, 519–528. doi: 10.1093/biomet/92.3.519

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Trans. Am. Math. Soc.* 54, 426. doi: 10.1090/S0002-9947-1943-0012401-3

Wu, J.-Y. (2017). The indirect relationship of media multitasking self-efficacy on learning performance within the personal learning environment: implications from the mechanism of perceived attention problems and self-regulation strategies. *Comput. Educ.* 106, 56–72. doi: 10.1016/j.compedu.2016.10.010

Wu, J.-Y., and Kwok, O. (2012). Using structural equation modeling to analyze complex survey data: a comparison between design-based single-level and model-based multi-level approaches. *Struct. Equat. Model. A Multidiscipl. J.* 19, 16–35. doi: 10.1080/10705511.2012.634703

Wu, J.-Y., Kwok, O., and Willson, V. L. (2014). Using design-based latent growth curve modeling with cluster-level predictor to address dependency. *J. Exp. Educ.* 82, 431–454. doi: 10.1080/00220973.2013.876226

Wu, J.-Y., Lin, J. J. H., Nian, M.-W., and Hsiao, Y.-C. (2017). A solution to modeling multilevel confirmatory factor analysis with data obtained from complex survey sampling to avoid conflated parameter estimates. *Front. Psychol.* 8:1464. doi: 10.3389/fpsyg.2017.01464

Yuan, K.-H., and Bentler, P. M. (2003). Eight test statistics for multilevel structural equation models. *Comput. Stat. Data Anal.* 44, 89–107. doi: 10.1016/S0167-9473(02)00349-3

Yuan, K.-H., and Bentler, P. M. (2007). Multilevel covariance structure analysis by fitting multiple single-level models. *Soc. Methodol.* 37, 53–82. doi: 10.1111/j.1467-9531.2007.00182.x

Zimprich, D., and Martin, M. (2009). "A multilevel factor analysis perspective on intellectual development in old age" in *Aging and Cognition: Research Methodologies and Empirical Advances*, eds H. B. Bosworth and C. Hertzog (Washington, DC, US: American Psychological Association), 53–76.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling editor is currently editing co-organizing a Research Topic with one of the author JW and confirms the absence of any other collaboration.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read
for greatest visibility
and readership

FAST PUBLICATION
Around 90 days
from submission
to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative,
and constructive
peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers
acknowledged by name
on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: info@frontiersin.org | +41 21 510 17 00

REPRODUCIBILITY OF RESEARCH
Support open data
and methods to enhance
research reproducibility

DIGITAL PUBLISHING
Articles designed
for optimal readership
across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics
track visibility across
digital media

EXTENSIVE PROMOTION
Marketing
and promotion
of impactful research

LOOP RESEARCH NETWORK
Our network
increases your
article's readership