THE RISK OF USING AN AVERAGE SCORE AS A LATENT VARIABLE IN MULTILEVEL MODELS

A Dissertation

by

CHI-NING CHANG

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Oi-Man Kwok
Committee Members,	Myeongsun Yoon
	Wen Luo
	Debra Fowler
	Lei-Shih Chen
Head of Department,	Shanna Hagan-Burke

May 2020

Major Subject: Educational Psychology

ABSTRACT

Educational researchers frequently work with data measured as multilevel structures; sometimes, they are also interested in latent constructs that cannot be directly observed and measured. Therefore, handling data dependency and measurement error issues is particularly important in statistical modeling. Multilevel Structural Equation Modeling (MSEM) is a promising approach to dealing with both issues. However, educational researchers still prefer Multi-Level Modeling (MLM) to MSEM. Conventional MLM cannot address the data dependency issue in within-level predictors. In addition, it cannot include a measurement model to handle measurement errors and construct a latent factor. As such, computing an average score to represent a latent factor in MLM is a common alternative approach in educational studies. This study evaluated the consequence of using an average score to represent a latent factor in MLM. The simulation results suggested that the bias of using an average score to represent a latent predictor in MLM is acceptable only when the following criterion are met: (1) group-mean centering or latent-mean centering is utilized; (2) the within-level factor loading of each item is equal to or above .80 (i.e., within-level composite reliability $\omega \geq 0.88$). Otherwise, MSEM is recommended.

DEDICATION

I dedicate this dissertation to my parents and my wife who give me unconditional love and support and encourage me to pursue my dream.

ACKNOWLEDGEMENTS

I would like to thank my committee chair, Dr. Oi-Man Kwok, for being a great mentor to me throughout my graduate years. I would also like to thank my committee members, Dr. Luo, Dr. Yoon, Dr. Fowler, and Dr. Chen for their guidance and support. I am very thankful for all faculty members and colleagues from the Research Measurement and Statistics (RMS) program, Center for Teaching Excellence (CTE), and Data-Enabled Discovery and Design of Energy Materials (D^3EM) program. I enjoyed working with them and learned a lot from them about being a multidimensional scholar. I wish to express my appreciation to my RMS friends, CTE colleagues, D^3EM team, and other friends in Texas, for making my time at Texas A&M University a wonderful experience. Finally, I owe many thanks to my wife, my parents, and Guan for the unconditional support.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by a dissertation committee consisting of Drs. Oi-Man Kwok, Wen Luo, Myeongsun Yoon, and Debra Fowler of the Department of Educational Psychology and Dr. Lei-Shih Chen of the Department of Health and Kinesiology. A consultation with Dr. Mark Lai was very helpful to understand the calculation of the variance of a composite score in MLM.

All work for the dissertation was completed independently by the student.

Funding Sources

Graduate study was supported by a research grant funded by the Ministry of Education, Taiwan.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER I

INTRODUCTION

Educational researchers frequently work with data measured as multilevel structures, for instance, student-level survey or demographic data, and school-level administrative or aggregate contextual data. As such, the use of data with multiple levels is common. Methodologists have already warned that when analyzing multilevel data with traditional linear models (e.g., multiple linear regression), the results (e.g., tests of significance) could be inaccurate due to the disregard of data dependency. For example, for a two-level dataset with students nested within schools, students from the same school are likely not to completely independent from each other. Without adequately taking this non-independent observation issue into account for the analyses, the results, especially the test of significance, can be biased and lead to incorrect statistical conclusions (Raudenbush & Bryk, 1986).

Multi-Level Modeling (MLM), also known as Hierarchical Linear Modeling (HLM; Raudenbush & Bryk, 1986), has become a widely used approach for analyzing multilevel data. This technique is a useful method to separate school effects from student and family inputs (Sellström & Bremberg, 2006). More specifically, researchers can identify how much of the variation in student outcomes is attributed to individual effort or family background at the student level (also called Level 1, the within level, or individual level) and how much is related to differences between schools (i.e., effects from the school level, also called Level 2, the between level, or cluster level). Nevertheless, MLM still has limitations. This study aims to evaluate the risk when researchers do not properly address the methodological issues in MLM.

In some situations, MLM may not work well. For example, examining the relationships between intrinsic motivational factors (e.g., science interest, math self- efficacy, and subjective task values) and student educational outcomes and career trajectories has been a long-standing interest of educational researchers (e.g., Schneider et al. 2016; Wang, 2013; Wigfield & Eccles, 2000). Given that MLM cannot include a measurement model for a motivational factor or a latent factor, researchers tend to theorize a construct, compute a composite score through a set of observed items from a scale, and utilize the composite score that is assumed to be free of measurement error in the analysis.

Computing an average score to represent a latent factor is a common approach in educational studies, especially for studies using the Trends in International Mathematics and Science Study (TIMSS) data established by the International Association for the Evaluation of Educational Achievement (IEA). TIMSS is an international dataset that provides policy makers and practitioners with insights for math and science education in the form of collected international assessments of knowledge and attitudes in math and science from students across 70 countries since 1995. With the need to analyze these latent measures, TIMSS (e.g., 2003, 2007) furnishes researchers with an average score for each non-cognitive measure (Martin & Preuschoff, 2008; Mullis, Martin, & Foy, 2008). Hence, researchers can conveniently utilize the composite scores representing latent constructs in their analyses.

However, when modeling latent factors by merely using a composite score, rather than the original items, two methodological issues could potentially bias the analytic results in MLM. The first issue is measurement errors. Since the items within a latent factor have been converted into a composite score, analysis without considering the measurement error of each item could lead to biased path coefficients (Hsiao, Kwok, & Lai, 2018; Rose, Wagner, Mayer, & Nagengast, 2019). Additionally, the regression family approaches (e.g., regression, MLM) are based upon the assumption that predictors are free of measurement error (Curran, 2003; Jaccard & Wan, 1995). Unless the observed composites are measured perfectly, the analytic results are biased. The second issue is the existence of data dependency in predictors. In MLM, only the variance of the outcome variable is separated into different levels. For predictors (especially the lower level or within-level predictors), their variances are assumed to be all from the same level. In other words, for the within-level predictors, the variations are all from the same within level. However, this assumption is too restrictive and, in many situations (e.g., educational studies), the variation of each within-level predictor may not solely come from "within" but may also include variation between clusters. The potential impact of ignoring the between-level variance for the within-level predictor has not yet been thoroughly examined.

Multilevel Structural Equation Modeling (MSEM) is an alternative way of analyzing multilevel observed data. The advantage of MSEM is the possibility to combine Structural Equation Modeling (SEM) with MLM; in other words, researchers can simultaneously estimate all the measurement errors and path parameters at different levels. The variance of each within-level variable (including predictors) can be separated into different levels, so we may estimate more authentic interrelationships among predictors and outcomes, partialling out the measurement errors at multiple levels (Preacher, Zhang, & Zyphur, 2011; Preacher, Zyphur, & Zhang, 2010).

Despite the ability of MSEM to possibly address the previously mentioned issues, educational researchers still prefer MLM to MSEM. As of August 1, 2019, the search results from the Web of Science database revealed that over the past few decades in educational research, 75 studies utilized MSEM, while 640 studies employed MLM or HLM. A similar trend was also found in TIMSS research. After further refining the search results to studies using the TIMSS datasets, the results showed that the studies employed MSEM only four times, whereas 23 TIMSS research projects utilized MLM (HLM).

Given the described risk from the failure to account for measurement errors and data dependency in within-level predictors, the purpose of this study was to evaluate the bias of estimating the relationship between an average composite score (representing a latent predictor) and a continuous outcome in MLM by comparing MLM results with MSEM results. A Monte Carlo study with 1,440 simulation factors was conducted. The simulation factors included the level of the intraclass correlation coefficients (ICC) for the predictor and outcome, the level of factor loadings of the latent predictor at the between- and within- levels, as well as multiple centering strategies. Considering the potential impact on math and science education policy and instructional decision-making around the world based on TIMSS research, this study also aimed to generate methodological insights for TIMSS researchers. The simulation followed the multilevel settings of TIMSS, such as cluster size = 30 and average number of clusters = 150. The results provide guidance on selecting adequate modeling strategies for a variety of complex scenarios.

CHAPTER II

LITERATURE REVIEW

Multilevel Structural Equation Modeling

Multilevel Structural Equation Modeling (MSEM) has been available for decades (e.g., Hox, 1995; McDonald & Goldstein, 1989; Muthén, 1989, 1994). With the recent progress of computer science and technology, the MSEM routine has been mostly available in SEM software such as M*plus*, LISREL, and Stata, allowing researchers to examine the relationships among latent and observed variables under multilevel data structures (Li & Beretvas, 2013).

In two-level data with observations nested within clusters (e.g., students nested within schools), the two sources of random variation are (a) random variation due to between-cluster differences at the between level and (b) random variation owing to differences among individuals within clusters at the within level. Assuming a balanced design, the data vector y_{ij}, a p-dimensional response vector with a total of N individuals (i) nested within J clusters ($i = 1 \ldots N$ individuals and $j = 1 \ldots J$ groups), can be decomposed into a within-level random component (y_{Wij}) and between-level random component (y_{Bj}) (Ryu, 2015):

$$y_{ij} = y_{Bj} + y_{Wij}\,, \tag{1}$$

where $E\,(y_{Bj}) = \mu_y$, $E\,(y_{Wij}) = 0$, $Cov\,(y_{Bj}\,,\, y_{Wij}) = 0$, and $E\,(y_{ij}) = \mu_y$. The within-level random component (y_{Wij}) and between-level random component (y_{Bj}) are uncorrelated and can be modeled (Hox, 2013; Ryu, 2015) by

$$y_{Wij} = \Lambda_W \eta_W + \varepsilon_W$$

$$y_{Bj} = \mu + \Lambda_B \eta_B + \varepsilon_B \tag{2}$$

By combining Equations (1) and (2), we obtain

$$y_{ij} = \mu + \Lambda_W \eta_W + \Lambda_B \eta_B + \varepsilon_B + \varepsilon_{W.} \tag{3}$$

where μ is a p-dimensional vector of grand means, and Λ_W is a $p \times m$ within-level factor-loading matrix, and m represents the number of within-level factors; η_W is a m-dimensional vector of within-level factor scores, and Λ_B is a $p \times h$ between-level factor loading matrix, where h shows the number of between-level factors; η_B is a h-dimensional vector of between-level factor scores; ε_B is a p-dimensional vector of between-level unique factors/measurement errors, while ε_W is a p-dimensional vector of within-level unique factors/measurement errors (Hsu, Lin, Kwok, Acosta, & Willson, 2016).

Measurement Model

Estimation of a measurement model begins with a partitioning of the total covariance matrix (Σ_T) into the between-level and within-level covariance matrices (i.e., Σ_B and Σ_W, respectively):

$$Cov(y_{ij}) = \Sigma_T = \Sigma_B + \Sigma_W \tag{4}$$

The between-level and within-level covariance matrices of the two-level confirmatory factor analysis (CFA) model can be expressed as follows:

$$\Sigma_B = \Lambda_B \Psi_B \Lambda'_B + \Theta_B$$

$$\Sigma_W = \Lambda_W \Psi_W \Lambda'_W + \Theta_{W}, \tag{5}$$

where Λ_B and Λ_W represent the factor-loading matrices for the between-level and within-level components, respectively; Ψ_B and Ψ_W are the factor covariance matrices for the between-level and within-level components, respectively; Θ_B and Θ_W are the covariance matrices of the unique factors (measurement errors) for the between-level and within-level components, respectively (Hsu et al., 2016).

Structural Model

The estimation of a structural model in MSEM is similar to the one in a conventional single-level SEM model. The model-implied covariance matrix $\hat{\Sigma}$ of a single-level SEM model for a set of *l* exogenous variables regressed on *k* exogenous latent variables and a set of *m* endogenous latent variables regressed on *n* endogenous latent variables with both measurement and structural models can be written as follows:

$$\hat{\Sigma} = \begin{bmatrix} \Lambda_Y\,(I-B)^{-1}(\Gamma\Phi\Gamma' + \Psi)[(I-B)^{-1}]'\Lambda'_Y + \Theta_\varepsilon & \Lambda_Y\,(I-B)^{-1}\Gamma\Phi\Lambda'_X \\ \Lambda_X\Phi\Gamma'[(I-B)^{-1}]'\Lambda'_Y & \Lambda_X\Phi\Lambda'_X + \Theta_\delta \end{bmatrix}, \tag{6}$$

where $\Lambda_X(l \times k)$ and Λ_Y $(m \times n)$ represent factor-loading matrices for exogenous variables *X* and endogenous variables *Y*, respectively; *B* represents a $n \times n$ square matrix containing the structural path coefficients from endogenous to other endogenous factors; Γ is a $n \times k$ matrix whose elements are the structural regression parameters from exogenous to endogenous factors; Φ and Ψ represent the $k \times k$ and $n \times n$ covariance matrices for exogenous factors ξ and the endogenous factors η, respectively; Θ_δ and Θ_ε represent the $l \times l$ and $m \times m$ covariance matrices for the measurement errors, δ and ε, respectively. In MSEM, Equation (6) can be extended to use the corresponding between-level and within-level components for measurement and structural models, representing the between-level and within-level matrices, $\hat{\Sigma}_B$ and $\hat{\Sigma}_y$, respectively (Li & Beretvas, 2013).

Parameter Estimation

A multilevel full information maximum likelihood (F_{ML}) estimation is commonly utilized for estimating parameters in multilevel models (Hsu et al., 2016; Ryu & West, 2009). Assuming multivariate normality for each level component and balanced case in which each cluster had equal individuals, the F_{ML} fitting function for the two-level structural equation model can be expressed as (Hsu et al., 2016):

$$F_{ML} = F_B\ (\theta) + F_W\ (\theta) = \sum_{j=1}^{J} \{tr[\Sigma_{SB}^{-1}(\theta)S_B] + log\,|\,\Sigma_{SB}(\theta)|\} + (N - J)\{tr[\Sigma_W^{-1}(\theta)S_W] + log\,|\,\Sigma_W(\theta)|\}\ , \quad (7)$$

where $F_B\ (\theta)$ and $F_W\ (\theta)$ are the between-level and within-level fitting functions; θ is the vector of the estimated parameters corresponding to a specified model; J denotes the number of clusters, while N is the cluster size; $\Sigma_{SB}(\theta)$ represents the implied between-level covariance matrix, and S_B is the between-level sample covariance matrix; $\Sigma_W(\theta)$ represents the implied within-level covariance matrix, and S_W is the within-level sample covariance matrix (Hsu et al., 2016).

Overall, the advantage of utilizing MSEM is the capability of combining SEM with MLM. First, to model latent factors, researchers can use measurement models to construct latent factors measured by a number of observed items, and the measurement models provide the entire model with important measurement information (e.g., factor loadings, measurement errors) for estimating less-biased parameters. Second, the covariance structure is partitioned into the between-level and within-level structures, which are not correlated with each other. The variations for both within-level variables, even for predictors, can be decomposed into two levels. The effects associated with the within-level variables are also partitioned into between- and within-levels; thus, researchers can investigate the multilevel effects. To sum up, both measurement error and data dependency issues are handled in MSEM.

Multi-Level Modeling

Multi-Level Modeling (MLM), also known as Hierarchical Linear Modeling (HLM; Raudenbush & Bryk, 1986), is a regression-based analysis that takes the multilevel data structure into account and is being more widely used in social science than MSEM. A simple two-level MLM model with one within-level predictor (X_{ij}) and one between-level predictor (W_j) can be written as follows:

$$
\begin{aligned}
Y_{ij} &= \beta_{0j} + \beta_{1j}X_{ij} + r_{ij} \\
\beta_{0j} &= \gamma_{00} + \gamma_{01}W_j + u_{0j} \\
\beta_{1j} &= \gamma_{10} + \gamma_{11}W_j + u_{1j}\,, \qquad (8)
\end{aligned}
$$

where the subscript j is for the clusters (j = 1…J), and the subscript i is for individuals (i = 1…nj); Y_{ij} denotes an outcome variable at the within level for the i^{th} individual in the j^{th} cluster; β_{0j} represents the y-intercept of the regression line for the j^{th} cluster, and β_{1j} is the slope of the regression line for the j^{th} cluster; r_{ij} means the random error associated with the response for the i^{th} individual in the j^{th} cluster; γ_{00} is the overall mean intercept adjusted for Wj, and γ_{01} refers to the regression coefficient associated with W*j* relative to the intercept; u_{0j} denotes the random effect of the j^{th} cluster on the intercept adjusted for W*j*; γ_{10} is the overall mean slope adjusted for W*j*, and γ_{11} refers to the regression coefficient associated with W*j* relative to the slope; u_{1j} denotes the random effect of the j^{th} cluster on the slope adjusted for W*j*.

Assumptions

The assumptions of the MLM model are shown below (Raudenbush & Bryk, 2002; Sullivan, Dukes, & Losina, 1999; Woltman, Feldstain, MacKay, & Rocchi, 2012):

$$
\begin{aligned}
&E(u_{0j}) = E(u_{1j}) = 0;\\
&E(\beta_{0j}) = \gamma_{00};\ E(\beta_{1j}) = \gamma_{10};\\
&var(\beta_{0j}) = var(u_{0j}) = \tau_{00};\ var(\beta_{1j}) = var(u_{1j}) = \tau_{11};\\
&cov(\beta_{0j}, \beta_{1j}) = cov(u_{0j}, u_{1j}) = \tau_{01};\ cov(u_{0j}, r_{ij}) = cov(u_{1j}, r_{ij}) = 0.
\end{aligned} \tag{9}
$$

The means of the random effects (u_{0j} and u_{1j}) are assumed to be zero. β_{0j} and β_{1j} have normal multivariate distributions with variances defined by τ_{00} and τ_{11}, respectively, and means equal to γ_{00} and γ_{11}, respectively. The covariance between β_{0j} and β_{1j} is τ_{01}, being identical to the covariance between u_{0j} and u_{1j}. Finally, the between-level random effects (u_{0j} and u_{1j}) and within-level random effect (r_{ij}) are uncorrelated with each other.

Two other assumptions in MLM could be limitations for researchers. First, the within-level predictor (X_{ij}) is assumed to be an observed variable with error-free measurement. Latent variables and measurement models are not allowed. An alternative approach in practice is to compute a composite score through a set of observed items to represent a latent factor. However, by doing this, the measurement error for each item would not be taken into account in the analysis. Second, the variance for an outcome variable y_{ij} could be decomposed into the within-level and between-level; yet, the variances for within-level predictors are assumed to stay at the within-level. This is problematic in MLM because for each within-level predictor, if the between-level variance exists, it will be added to the within-level variance. Therefore, MLM may fail to deal with both data dependency and measurement error issues.

This is best illustrated with an example. Figure 1 shows an MSEM model where the between-level and within-level factors (η_B and η_W, respectively) can be measured by the observed Items 1 to 4 ($I1_{ij}$ to $I4_{ij}$), and the effect between the latent factor and outcome Y_{ij} at each level would be estimated. If one analyzing the same dataset computes an average (M_{ij}) from $I1_{ij}$ to $I4_{ij}$ to represent a within-level latent factor (M_w) in MLM (as shown in Figure 2), three things must be noticed: (a) the total variance of M_{ij} will be assumed to stay at the within-level (i.e., M_w); (b) the variance of this composite variable (M_{ij}) at the within level will be forced to include the between-level total variance, as shown in Equation (10); and (c) the measurement error variance mixed in the total variance cannot be partialled out. The variance of this composite variable is formally defined as follows:

$$\Sigma_{\mathrm{M}} = (\Sigma_{\mathrm{B}} + \Sigma_{\mathrm{W}}) / \mathrm{I}^2 , \tag{10}$$

where Σ_M represents the total variance of the within-level composite variable computed through a mean score approach; Σ_B and Σ_W are computed based on Equation (5) consisting of elements like factor loadings, factor variances (and covariance), as well as measurement error variances (and covariance); I denotes the number of observed items.

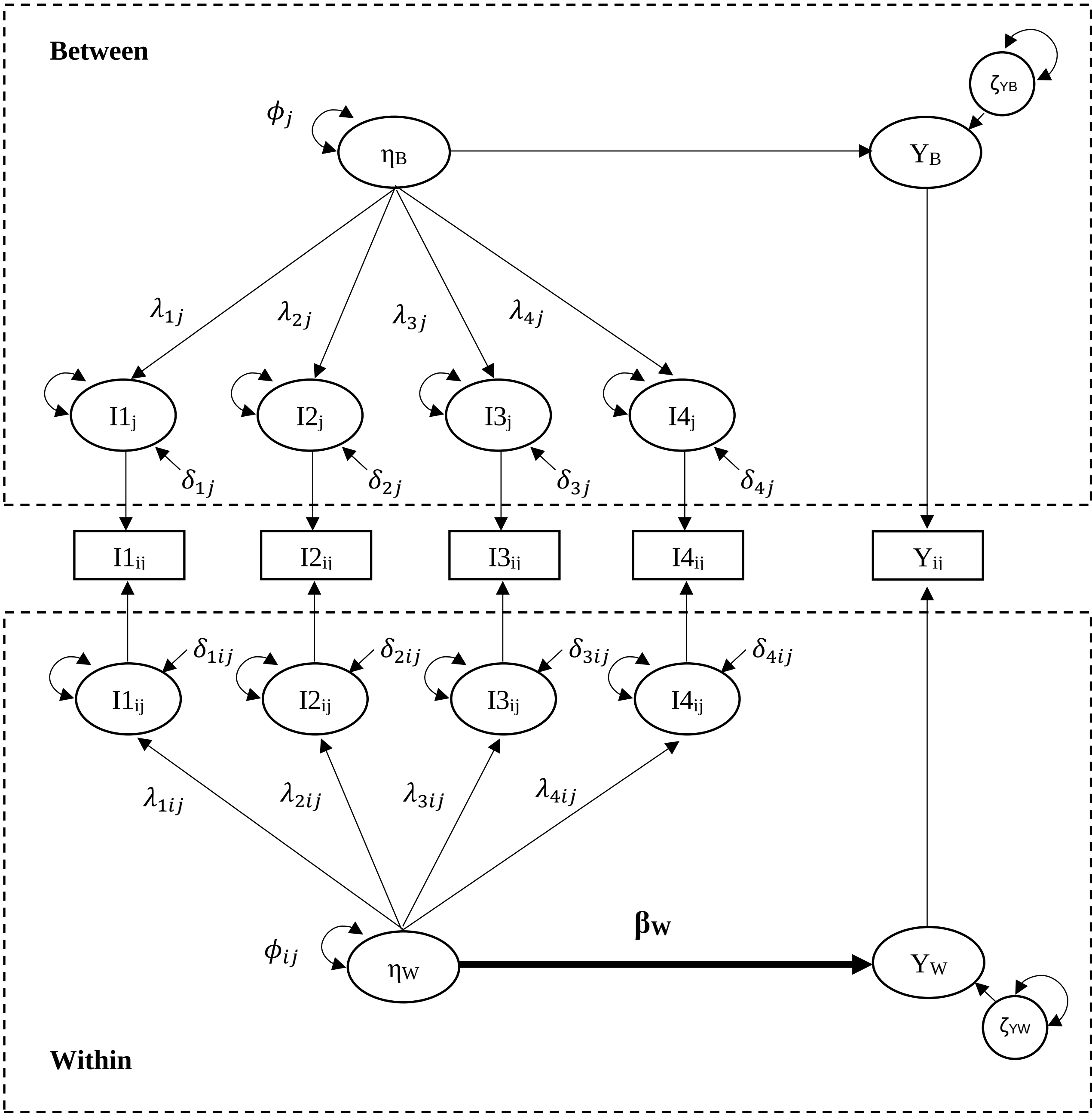

Figure 1. An MSEM example.

Given that the variance components of a composite score are mixed ($\Sigma_B + \Sigma_W$), we should be aware that in some situations, data dependency and measurement errors may cancel out. For example, in theory, a higher measurement error variance at the within level will lead to a lower total variance ($\Sigma_B + \Sigma_W$); at the same time, a higher between-level variance will result in a higher total variance. Hence, the variance components could counterbalance each other. This example demonstrates that using a composite score in MLM involves very complicated methodological issues, leading to an unreliable, risky result.

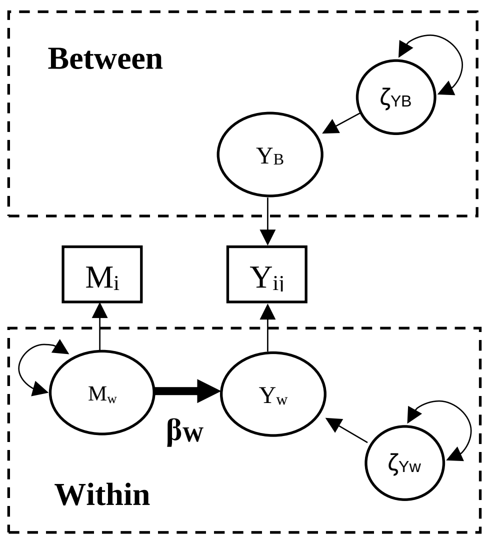

Figure 2. An MLM model for the current study.

Estimation of the Standardized Path Coefficient

When examining the effect between M_w and Y_w in MLM, researchers often report the standardized path coefficient (β_W). In Figure 2, β_W can be calculated as:

$$
\begin{aligned}
\beta w &= bw * sqrt\ (VAR_M / VAR_{YW}) \\
&= (COV(M,Yw) / VAR_M) * sqrt(VAR_M / VAR_{YW}) \\
&= (COV(M,Yw) / SD_M^2) * (SD_M / SD_{YW}) \\
&= COV(M,Yw) / (SD_M * SD_{YW}) \\
&= r_w * SD_M * SD_{YW} / (SD_M * SD_{YW}) = r_w \,,
\end{aligned}
\tag{11}
$$

where bw is an unstandardized regression coefficient computed through $COV(M,Yw) / VAR_M$; and COV means covariance; VAR_M represents the total variance of M_{ij} (i.e., Σ_M); VAR_{YW} is the variance of Y_{ij} at the within level; *sqrt* denotes square root; *SD* represents standard deviation; r_w means the correlation coefficient between M_w and Y_w; in this case, βw will be equal to r_w eventually.

Equation (11) reveals that the variance of the composite variable (VAR_M) shows a strong impact on the estimation of the effect between the composite variable (M_{ij}) and the outcome (Y_{ij}). In other words, to avoid obtaining a biased standard coefficient, the predictor's variance should be estimated accurately. However, as mentioned previously, in MLM, the variance estimation for each within-level variable is often inaccurate when the variable contains either measurement errors or between-level variance.

Centering Strategy

Two centering strategies in MLM can potentially deal with the data dependency of within-level predictors. Both Raudenbush and Bryk (2002) and Enders and Tofighi (2007) suggested using group-mean centering ($x_{ij} - \bar{x}_{.j}$) to decompose the within-level predictor into the between-level and within-level effects. A significant drawback in group-mean centering is that the observed group mean ($\bar{x}_{.j}$) may contain measurement errors. Several questions arise: (1) are the samples from each cluster randomly and sufficiently selected? and (2) are missing data completely at random? If not, centering observed group mean is inaccurate because the observed group mean is not identical to the true group mean (Shin & Raudenbush, 2010).

One promising approach is to utilize Latent-Mean Centering in MLM (LMC-MLM) (Asparouhov & Muthén, 2019). When estimating the relationship between an observed variable (X_{ij}) and a continuous outcome (Y_{ij}), LMC-MLM (without a random slope) can be described as in the following equations:

$$
\begin{aligned}
&X_{ij} = X_{W,ij} + X_{B,j} \\
&Y_{ij} = \alpha_j + \beta_1 X_{W,ij} + \varepsilon_{W,ij} \\
&\alpha_j = \alpha + \beta_2 X_{B,j} + \varepsilon_{B,j} \\
&\varepsilon_{W,ij} \sim N(0, \sigma_W),\ \varepsilon_{B,j} \sim N(0, \sigma_B),\ X_{W,ij} \sim N(0, \psi_W),\ X_{B,j} \sim N(\mu, \psi_B).
\end{aligned} \tag{12}
$$

The main difference from the conventional group-mean centered MLM is the identification of the latent group mean ($X_{B,j}$) for each group, which is an unknown value that can be estimated to account for the sampling error in the mean estimate through Bayesian estimation algorithms (Asparouhov & Muthén, 2019). However, even though LMC-MLM seems more promising, the observed variable (X_{ij}) is still assumed to be free of measurement error under the MLM theoretical framework. If X_{ij} is a composite score containing measurement errors, there is less evidence in the literature to show whether and to what extent the latent-mean centering can improve biased estimates in this particular case.

TIMSS Research

In practice, researchers generally utilize observed composites in their studies (Hsiao et al., 2018), especially in the Trends in International Mathematics and Science Study (TIMSS) research. The TIMSS consists of both an international large-scale survey and assessments conducted by the International Association for the Evaluation of Educational Achievement (IEA) that monitor trends in students' math and science in Grades 4 and 8 across 70 countries since 1995 (IEA TIMSS & PIRLS International Study Center, 2019). Given that the TIMSS datasets collect numerous variables related to student achievement and motivational factors in math and science, many researchers have engaged in analyzing their country's data within this dataset to investigate the relationships among student ability, school average ability, and student academic self-concept, as in examinations of the big-fish-little-pond effect (BFLPE) (Marsh & Parker, 1984).

As expected, many BFLPE studies constructed academic self-concept in MLM by using a composite score. Reviewing the articles published in high impact journals (2001-2018) indexed by the Web of Science, I found that more than half of the BFLPE studies (16 out of 29) noticed the multilevel structure of the TIMSS data, as these studies utilized MLM or MSEM. As shown in Table 1, among these 16 studies, only four demonstrated explicit awareness of the measurement error issue by employing measurement models to construct latent factors (e.g., math/science self-concept) in the MSEM models. The other 12 studies using MLM relied on composite scores to represent latent factors. Interestingly, most of these 12 studies directly used average scores as latent factors because TIMSS (e.g, 2003, 2007) had already computed an average for each non-cognitive measure in the released datasets (Martin & Preuschoff, 2008; Mullis, Martin, & Foy, 2008). Regarding the centering strategies, three studies did not report on this, two studies used group-mean centering, and the other seven studies employed grand-mean centering. Yet, all 12 studies using MLM failed to provide a rationale about a decision in the choice of centering strategy. Theoretically, as previously mentioned, MLM has limitations in modeling latent factors. However, TIMSS researchers still tend to use the MLM approach, and their studies have been cited by other studies. The number of citations range from six to 215.

Table 1

TIMSS Research (2001-2018) Using MSEM or MLM as Indexed by the Web of Science

Author (Year)	Journal	Country	Data	Method	Composite	Centering	Citations
Guo, Marsh, Parker, & Dicke (2018)	Learning and Instruction	15 OECD countries	TIMSS and PIRLS 2011	MSEM			3
Wang & Bergin (2017)	Learning and Individual Differences	59 countries and regions	TIMSS 2011	MSEM			3
Wang (2015)	Educational Psychology	49 countries	TIMSS 2007	MSEM			13
Marsh et al. (2014)	Journal of Cross-Cultural Psychology	US and Saudi Arabian	TIMSS 2007	MSEM			40
Liou & Jessie (2018)	Research Papers in Education	Taiwan	TIMSS 2007	MLM	Average	Grand mean	8
Wang & Liou (2017)	International Journal of Science Education	Taiwan	TIMSS 2011	MLM	IRT	Group mean	13
Min, Cortina, & Miller (2016)	Learning and Individual Differences	13 countries	TIMSS 2003, 2007 and 2011	MLM	Average	Unknown	8

Table 1 (continued)

Author (Year)	Journal	Country	Data	Method	Composite	Centering	Citations
Sheldrake (2016)	Learning and Individual Differences	England	TIMSS 2011	MLM	IRT	Unknown	6
Tsai & Yang (2015)	International Journal of Science Education	Taiwan	TIMSS 2011	MLM	Average	Unknown	14
Liou (2014a)	International Journal of Science Education	Taiwan	TIMSS 2003 and 2007	MLM	Average	Grand mean	11
Liou (2014b)	The Asia-Pacific Education Researcher	Taiwan	TIMSS 2011	MLM	Average	Grand mean	12
Mohammadpour & Abdul Ghafar (2014)	Scandinavian Journal of Educational Research	48 countries	TIMSS 2007	MLM	Average	Grand mean	17

Table 1 (continued)

Author (Year)	Journal	Country	Data	Method	Composite	Centering	Citations
Mohammadpour (2013).	Learning and Individual Differences	Singapore	TIMSS 2007	MLM	Average	Grand mean	31
Mohammadpour (2012a).	Science Education	Malaysia	TIMSS 1999, 2003, and 2007	MLM	Average	Grand mean	24
Mohammadpour (2012b).	The Asia-Pacific Education Researcher	Singapore	TIMSS 2007	MLM	Average	Grand mean	25
Wilkins (2004)	The Journal of Experimental Education	41 countries	TIMSS 1999	MLM	Average	Group mean	215

Note. Literature search date: Feb. 8, 2019. The number of citations was counted on Aug. 6, 2019, as provided by Google Scholar.

Gaps in the Literature and Purpose of the Study

To the best of my knowledge, when using an average score as a latent predictor in MLM, the performance of estimating the relationship between the latent predictor and outcome has yet to be investigated—not to mention when a centering strategy is also involved. Furthermore, the TIMSS often attracts worldwide attention when announcing world rankings for student math and science achievement of each country. The impact of TIMSS research might ripple through education policy and curriculum decisions in each country. Therefore, a methodological study for using composite scores in MLM is needed.

The present study aimed to conduct a simulation study to evaluate the risk of estimating the relationship between an average composite score (representing a latent predictor) and a continuous outcome in MLM. Given that MSEM properly handles both data dependency and measurement error issues, the MLM simulation results are used for comparison with MSEM results. The discrepancy between them would be considered as a bias, since the variance components show impacts in Equation (11). Given one must also consider the elements in Equation (5), the simulation factors included the level of the intraclass correlation coefficients (ICC) for the predictor and outcome, the level of factor loadings of the latent predictor at the between- and within-levels, as well as the multiple centering strategies. The simulation followed the multilevel settings of TIMSS (i.e., the cluster size of TIMSS is about 30, and the average number of clusters is approximately 150). The number of items for a latent factor was set to four based on the number of items for math/science self-concept in TIMSS (Liou, 2014a, 2014b). The results will then provide guidance on selecting adequate modeling strategies under a variety of complex scenarios.

CHAPTER III

METHODS

Data Generation

Figure 3 was used as the population model for generating simulation datasets. The population model consisted of a measurement model and an outcome variable at the between level and within level. In the measurement model, four observed items ($I1_{ij}$ *to* $I4_{ij}$) were partitioned into the between level and within level. The partitioned variances were loaded on the between latent factor (η_B) and within latent factor (η_W), respectively. b_B and b_W denoted the effects of the latent factor on the outcome (Y_{ij}) at the between level and within level, respectively. The residual variance of Y_{ij} was partitioned into the between level and within level (i.e., ζ_{YB} and ζ_{YW}, respectively).

At the within level, four observed items were loaded on η_W. The factor variance η_W was set at 1.0. Six sets of factor loadings (λ_{1ij} to λ_{4ij}) were set to (.5,.5,.5,.5), (.6,.6,.6,.6), (.7,.7,.7,.7), (.8,.8,.8,.8), (.9,.9,.9,.9), and (.99,.99,.99,.99), respectively; the corresponding measurement errors (δ_{1ij} to δ_{4ij}) were (.75,.75,.75,.75), (.64,.64,.64,.64), (.51,.51,.51,.51), (.36,.36,.36,.36), (.19,.19,.19,.19), and (.02,.02,.02,.02), respectively. b_W was set to .50. The residual variance ζ_{YW} was 1.0. All the parameters at the within level are summarized in Table 2.

Table 2

Population Model Parameters at the Within Level

η_W	ζ_{YW}	b_W	λ_{1j} to λ_{4j}	δ_{1j} to δ_{4j}
1.0	1.0	.50	.5,.5,.5,.5	.75,.75,.75,.75
1.0	1.0	.50	.6,.6,.6,.6	.64,.64,.64,.64
1.0	1.0	.50	.7,.7,.7,.7	.51,.51,.51,.51
1.0	1.0	.50	.8,.8,.8,.8	.36,.36,.36,.36
1.0	1.0	.50	.9,.9,.9,.9	.19,.19,.19,.19
1.0	1.0	.50	.99,.99,.99,.99	.02,.02,.02,.02

The between-level model had an identical structure to the within-level model. The parameters of the between-level population model are listed in Tables 3 and 4. b_B was set to .50. To create different intraclass correlation coefficient (ICC) conditions (i.e., .01, .10, .30, .50) for outcome Y_{ij}, the residual variance ζ_{YB} was set to .01, .11, .43, and 1.00 based on the following equation:

$$ICC = \frac{\tau_{00}}{\tau_{00}+\sigma^2} \ , \tag{13}$$

where τ_{00} is the between-level variance for Y_{ij} , and σ^2 is the within-level variance for Y_{ij}. In other words, *ICC* represents the proportion of the total variance of Y_{ij} that is accounted for by the between level.

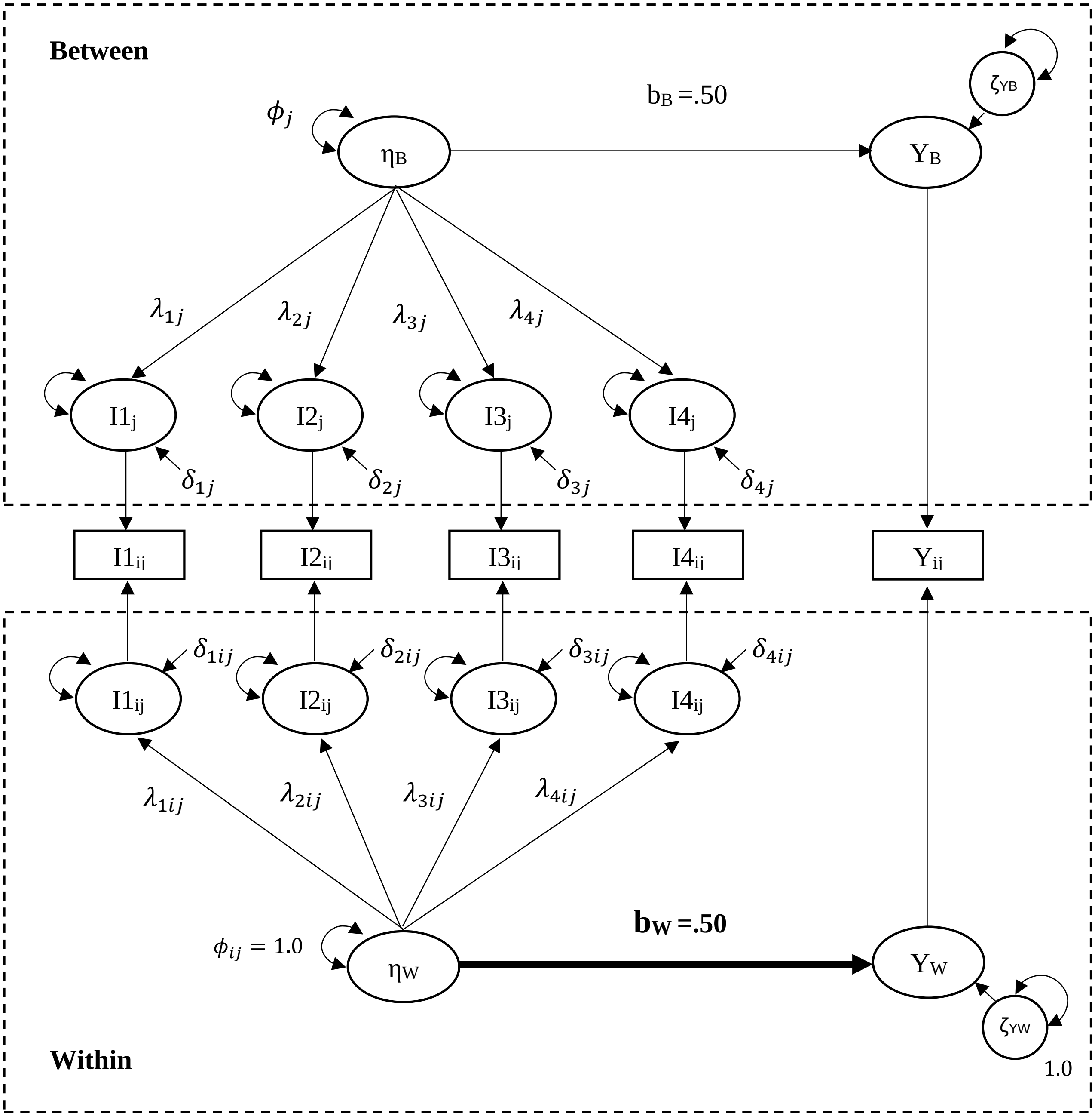

Figure 3. The population model for generating simulation datasets.

Table 3

Residual Variance at the Between Level for Outcome Y_{ij}

ζ_{YB}	b_B
.01 (ICC$_y$=.01)	.50
.11 (ICC$_y$=.10)	.50
.43 (ICC$_y$=.30)	.50
1.0 (ICC$_y$=.50)	.50

In the same vein, to create different latent factor ICC conditions (i.e., .01, .10, .30, .50), the factor variance η_B was also set to .01, .11, .43, and 1.00, respectively. Three sets of standardized factor loadings were (.5,.5,.5,.5), (.7,.7,.7,.7), and (.99,.99,.99,.99), and the corresponding measurement errors (δ_{1j} to δ_{4j}) were (.75,.75,.75,.75), (.51,.51,.51,.51), and (.02,.02,.02,.02), respectively. However, in some conditions, η_B was not equal to 1.0. The unstandardized between-level factor loadings (λ_{1j} to λ_{4j}) had to be recalculated through the following equation to keep the identical between-level variance (Σ_B) in Equation (5):

$$\lambda = \text{sqrt}\ (\ \lambda^2_{\text{standardized}}\ /\ \eta_B). \tag{14}$$

The corresponding λ_{1j} to λ_{4j} and δ_{1j} to δ_{4j} are shown in Table 4.

The study applied the Monte Carlo procedure in M*plus* 8 (Muthén & Muthén, 1998-2017). Tables 2–4 summarize the 288 (*6 × 4 × 12*) conditions for the population model. For each condition, 500 datasets were generated. Following the multilevel settings of the TIMSS design, each simulation dataset contains 150 clusters with cluster size = 30. Datasets were created based on a standard multivariate normal distribution utilizing a randomly chosen seed. The *MLR* was applied to obtain the model solutions.

Table 4

Population Measurement Model at the Between Level for the Latent Predictor

η_B	$\lambda_{standardized}$	λ_{1j} to λ_{4j}	δ_{1j} to δ_{4j}
.01 (ICC_X=.01)	.5	5, 5, 5, 5	.75, .75, .75, .75
.01 (ICC_X=.01)	.7	7, 7, 7, 7	.51, .51, .51, .51
.01 (ICC_X=.01)	.99	9.9, 9.9, 9.9, 9.9	.02, .02, .02, .02
.11 (ICC_X=.10)	.5	1.51, 1.51, 1.51, 1.51	.75, .75, .75, .75
.11 (ICC_X=.10)	.7	2.11, 2.11, 2.11, 2.11	.51, .51, .51, .51
.11 (ICC_X=.10)	.99	2.98, 2.98, 2.98, 2.98	.02, .02, .02, .02
.43 (ICC_X=.30)	.5	0.76, 0.76, 0.76, 0.76	.75, .75, .75, .75
.43 (ICC_X=.30)	.7	1.07, 1.07, 1.07. 1.07	.51, .51, .51, .51
.43 (ICC_X=.30)	.99	1.51, 1.51, 1.51, 1.51	.02, .02, .02, .02
1.0 (ICC_X=.50)	.5	.5, .5, .5, .5	.75, .75, .75, .75
1.0 (ICC_X=.50)	.7	.7, .7, .7, .7	.51, .51, .51, .51
1.0 (ICC_X=.50)	.99	.99, .99, .99, .99	.02, .02, .02, .02

Simulation Design Factors

Five design factors were considered in this study. These were (1) latent predictor ICC_X, (2) ICC_Y, (3) factor loadings at the within-level (λ_{1ij} to λ_{4ij}), (4) factor loadings at the between-level (λ_{1j} to λ_{4j}), and (5) misspecification types.

(1) Latent predictor ICC_X

The four levels of latent predictor ICC_X were set to .01, .10, .30, and .50. An ICC_X of .01 implies that there was no data dependency for the latent predictor. Note that the calculations of

ICC for the latent predictor and for the observed items within the measurement model were not identical. More details can be found in the study of Hsu et al. (2016).

(2) ICC_Y

The four levels of ICC_Y were set to .01, .10, .30, and .50.

(3) Factor loadings at the within level (λ_{1ij} to λ_{4ij})

Given that my primary focus was the within-level latent predictor, I set six levels of factor loadings for the within level (more than the between-level). The six levels of factor loadings at the within level were .50, .60, .70, .80, .90, and .99. The loading of .99 implied that the latent predictor was free of measurement error at the within-level.

(4) Factor loadings at the between level (λ_{1j} to λ_{4j})

The three levels of factor loadings at the between level were .50, .70, and .99. The loading of .99 implied that the latent predictor was free of measurement error at the between level.

(5) Misspecification types

Five misspecification types were considered, including the MSEM, uncentered MLM, grand-mean centered MLM, group-mean centered MLM, and latent-mean centered MLM models. All the MLM models were with random intercepts and no random slope. Given that MLM is unable to incorporate any measurement models, an average score was computed through observed items ($I1_{ij}$ to $I4_{ij}$) to represent a latent factor in all the MLM models, as shown in Figure 2. The latent-mean centering in the current study followed the study of Asparouhov and Muthén (2019) and Example 9.1 of the M*plus* 8 User's Guide (Muthén & Muthén, 1998-2017).

In sum, for each data-generating model, a 4 (latent predictor ICC_X) × 4 (ICC_Y) × 6 (factor loadings at the within level) × 3 (factor loadings at the between level) × 5 (misspecification type) factorial design was used, totaling 1,440 conditions.

Analysis of Simulation Results

The current study aimed to evaluate whether and to what extent using an average score in MLM leads to biased estimation. The primary focus was the performance of the standardized path parameter at the within level (β_W) across conditions. R packages (e.g., *MplusAutomation*, *Tidyverse*, etc.) and Microsoft Excel were used to analyze and produce visuals of the simulation results (Hallquist & Wiley, 2018). A few analyses of the simulation results were conducted.

First, as discussed in Chapter II, variance plays a key role in estimating β_W, as shown in Equation (11). This study explored the relationship between the variance of the average score (Var_M) and β_W in MLM by conducting a regression analysis, controlling for the variance of Y_{ij} at the within level (VAR_{YW}) and the unstandardized path coefficient (b_W). The results highlight the consequence of failing to partial out the error variance and between-level variance, providing insights in interpreting the evaluation results.

Second, to evaluate the performance of β_W in MLM, this study used the β_W of MLM models (i.e., uncentered, grand-mean centered, group-mean centered, and latent-mean centered MLM models) to compare with the β_W of MSEM at each condition (i.e., the same ICC_X, ICC_Y, factor loadings at the within level, and factor loadings at the between level). In other words, the β_W of MSEM was treated as a true population value at each condition. The major evaluation criteria for β_W were the (1) relative parameter bias, (2) relative standard error bias, and (3) root mean squared error (*RMSE*).

(1) Relative parameter bias

The relative parameter bias *RPB(θ)* was calculated as follows:

$$RPB(\theta) = R^{-1}\sum_{r=1}^{R}\frac{\hat{\theta}_r-\theta}{\theta} \ , \tag{15}$$

where $\hat{\theta}_r$ is the parameter estimate for replication *r*, *θ* stands for the population parameter, and *R* is the total number of replications. The acceptable *RPB* should be between -10% and 10% (Muthén & Muthén, 2002).

(2) Relative standard error bias

In a similar way, the relative standard error bias *RSEB(θ)* was calculated as follows:

$$RSEB(\theta) = R^{-1}\sum_{r=1}^{R}\frac{\widehat{M_{SE}}_r-SD}{SD} \ , \tag{16}$$

where $\widehat{M_{SE}}_r$ is the average of the estimated standard errors of the parameter estimate for replication *r*; *SD* stands for the true population value of this parameter, the standard deviation of the parameter estimate over the replications of the Monte Carlo study; *R* is the total number of replications. The acceptable *RSEB* should be between -10% and 10% (Muthén & Muthén, 2002).

(3) *RMSE*

The root mean squared error (*RMSE*) is a measure of overall accuracy (Ma, Raina, Beyene, & Thabane, 2012). The *RMSE* is defined as the square root of the sum of the variance and squared bias of the parameter estimate as follows:

$$RMSE = sqrt\ (Bias^2 + VAR(\theta)) \ , \tag{17}$$

where *Bias* represents $\hat{\theta}_r - \theta$ in Equation (15); $VAR(\theta)$ denotes the variance of the parameter estimate over the replications of the Monte Carlo study. The lower the *RMSE*, the higher the accuracy.

CHAPTER IV

RESULTS

The 4 (latent predictor ICC_X: .01, .10, .30, and .50) × 4 (ICC_Y: .01, .10, .30, and .50) × 6 (within-level factor loadings: .50, .60, .70, .80, .90, and .99) × 3 (between-level factor loadings: .50, .70, .99) × 5 (misspecification types: MSEM, uncentered MLM, grand-mean centered MLM, group-mean centered MLM, and latent-mean centered MLM models) factorial design yielded 1,440 simulation settings. As shown in Table 5, MSEM was the population model for each simulation condition (N = 288), while MLM models with centering approaches were evaluated whether and to what extent computing an average score for a latent construct in MLM leads to a biased standardized path coefficient. Given that the analytic results of the grand-mean centering and uncentered approach in MLM were identical, this study only presents and evaluates the results for grand-mean centering, group-mean centering, and latent-mean centering under different simulation conditions (N = 288 × 3).

Table 5

Simulation Settings for the Population Model and Misspecification Types

Population model	Misspecification types		
MSEM:	Grand-mean centered MLM:	Group-mean centered MLM:	Latent-meancentered MLM:
4 (latent predictor ICC_X) × 4 (ICC_Y) × 6 (within-level factor loadings) × 3 (between-level factor loadings) = 288 simulation conditions	4 (latent predictor ICC_X) × 4 (ICC_Y) × 6 (within-level factor loadings) × 3 (between-level factor loadings) = 288 simulation conditions	4 (latent predictor ICC_X) × 4 (ICC_Y) × 6 (within-level factor loadings) × 3 (between-level factor loadings) = 288 simulation conditions	4 (latent predictor ICC_X) × 4 (ICC_Y) × 6 (within-level factor loadings) × 3 (between-level factor loadings) = 288 simulation conditions

Note. The analytic results of grand-mean centered and uncentered MLM were identical.

The Relationship Between VAR_M and β_W in MLM

Equation (11) implies that the variance of the average composite score (VAR_M) shows a strong impact on the estimation of the effect (i.e., the within-level standardized path coefficient, β_W) between the within-level composite predictor (an average score computed through four items) (M_{ij}) and the outcome (Y_{ij}). To verify the relationship between VAR_M and β_W, I analyzed the simulation results across 864 conditions (i.e., 288 conditions × 3 centering approaches) in MLM. Variables collected from the simulation results included β_W, VAR_M, the within-level variance of the outcome Y_{ij} (VAR_{YW}), and the within-level unstandardized path coefficient (b_W), which are the important components of Equation (11).

The correlation matrix (Table 6) indicated the VAR_M and β_W are highly correlated (r = .945) without controlling for VAR_{YW} and b_W. I further conducted a multiple linear regression analysis to control for VAR_{YW} and b_W. The regression results (Table 7) showed that a one standard deviation increase in VAR_M was associated with a 1.461 standard deviation increase in β_W, over and above the VAR_{YW} and b_W; namely, the higher the VAR_M, the higher β_W. In other words, to avoid obtaining a biased β_W, VAR_M should be estimated accurately.

Table 6

Correlations (r) for VAR_M, VAR_{YW}, b_W, and β_W (N = 864)

	1	2	3	4
1. VAR_M	–			
2. VAR_{YW}	.820	–		
3. b_W	-.708	-.257	–	
4. β_W	.945	.829	-.540	–

Table 7

Results of the Multiple Linear Regression Analysis (N = 864)

	β	SE
VAR_M	1.461***	(.005)
VAR_{YW}	-.259***	(.019)
b_W	.428***	(.036)
Intercept	.067***	(.014)
R^2	.935***	

Note. Dependent variable = β_W; N = sample size; β = standardized coefficient; SE = standard error; *** p < .001.

Evaluating the Performance of β_W of MLM

Issues in examining the relationship (β_W) between a within-level average composite score (representing a latent predictor) and a continuous outcome in MLM include the measurement error estimation and data dependency of the within-level predictors. These issues might bias the estimation of the variance of the average composite predictor (VAR_M) and further lead to an inaccurate β_W.

Hence, to evaluate the performance of β_W in MLM, the β_W of MSEM was treated as a true population value at each simulation condition given its capability to handle both measurement error and data dependency issues. The evaluation criteria were the (1) relative parameter bias, (2) relative standard error bias, and (3) root mean squared error (*RMSE*). For each criterion, factors potentially influencing the estimation of VAR_M, such as centering strategies, the level of within-level factor loadings, the level of predictor's ICC (ICC_X), and the level of between-level factor loadings, were considered when evaluating the performance of β_W in MLM.

Relative Parameter Bias

The distribution of relative parameter bias values of estimating the β_W across conditions for each centering strategy in MLM by the level of within-level factor loadings is illustrated in boxplots in Figure 4. A range between two blue dashed lines (i.e., between -0.1 and 0.1) indicates an acceptable level of relative parameter bias; however, outside of this range, the β_W would be considered biased. Overall, when the within-level factor loading for each item was equal to or above 0.80, the relative parameter bias values for the group-mean centering and latent-mean centering across conditions were acceptable.

As discussed in Chapter II, group-mean centering and latent-mean centering have the capacity to partition the variance into the between level and within level. Only the within-level variance of VAR_M is used for estimating the within-level effect (β_W), no matter what the between-level factor loadings and predictor's ICC (ICC_X) are. As expected, in Figure 4, in each level of within-level factor loadings, there were no variations of relative parameter bias values across other between-level related conditions (ICC_Y, ICC_X, and between-level factor loadings) for group-mean centering and latent-mean centering.

Even though group-mean centered MLM and latent-mean centered MLM can handle the data dependency issue for the within-level predictors, these two approaches are not able to deal with the measurement error issue. As shown in Figure 4, when the within-level factor loading was .99 (almost perfect reliability with very little measurement error), the relative parameter bias values for group-mean centered MLM and latent-mean centered MLM were close to zero. However, the relative parameter bias became worse as the level of within-level factor loadings decreased (i.e., measurement error increased) from .99 to .50. Based on Equations (5) and (10) and previous regression results, the relative parameter bias became worse because the lower within-level factor loadings led to a smaller variance of the average composite predictor (VAR_M) and further resulted in an underestimated within-level standardized path coefficient (β_W) in MLM.

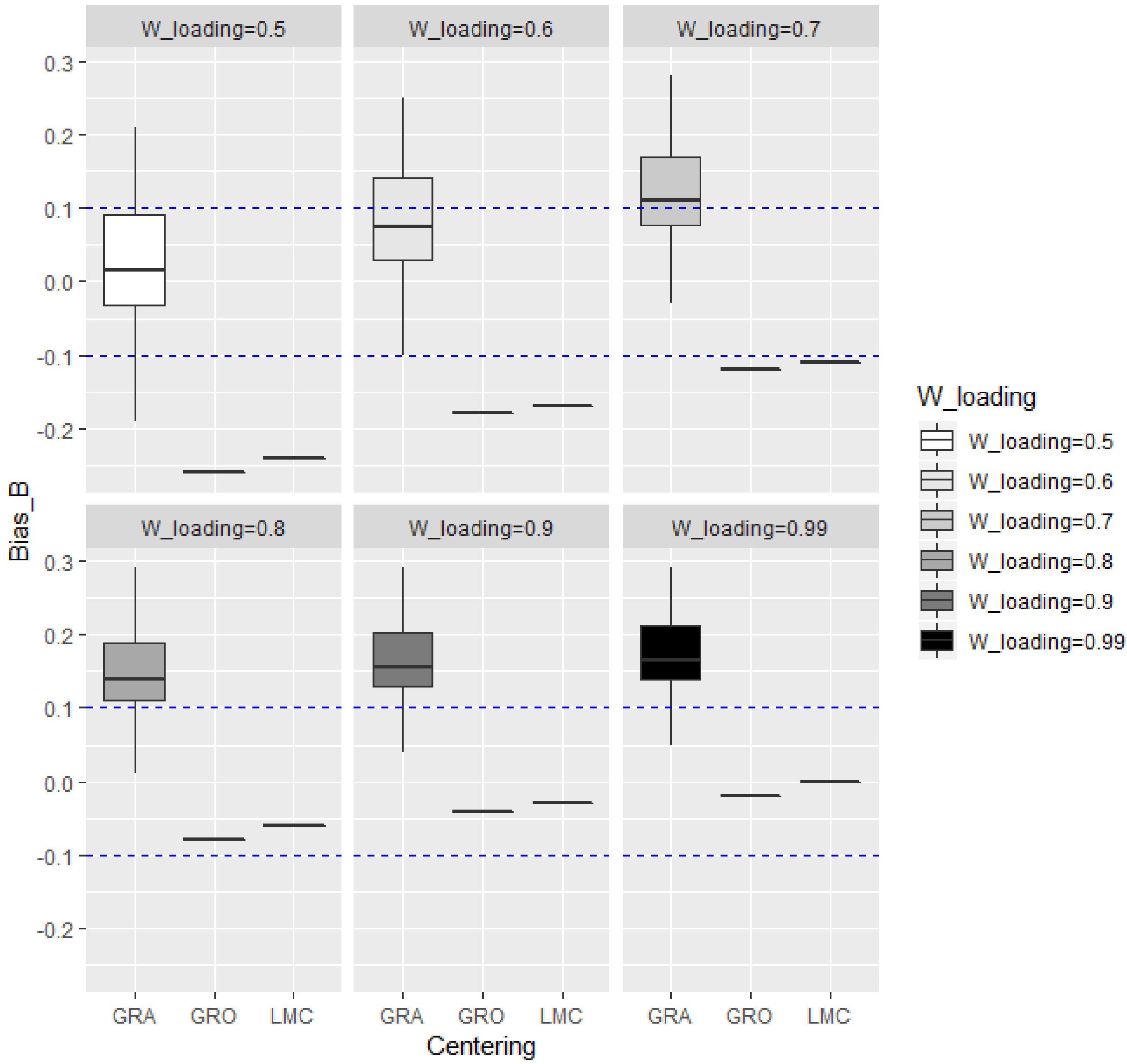

Figure 4. Boxplots showing the distribution of relative parameter bias values across conditions for each centering strategy in MLM by the level of within-level factor loadings.

Note. Bias_B = relative parameter bias of β_W; W_loading = within-level factor loading; GRA = grand-mean centering; GRO = group-mean centering; LMC = latent-mean centering.

Grand-mean centering cannot deal with both measurement errors and the data dependency of the within-level predictors. As shown in Figure 4, when the within-level factor loadings were .99 (almost perfect reliability with very little measurement error), most of the relative parameter bias values for grand-mean centered MLM were higher than 0.1 (out of the acceptable range). It was because the between-level variance was not separated out from VAR_M. The inflated VAR_M led to an overestimated β_W. Interestingly, as the within-level factor loading decreased from .99 to .50, the relative parameter bias became better but was still risky because the between-level variance and measurement error variance canceled out each other (as discussed in Chapter II). Therefore, grand-mean centered MLM is not recommended.

I further examined different levels of ICC_X as shown in Figure 5. As expected, the relative parameter bias values were identical across different conditions of ICC_X for group-mean centered MLM and latent-mean centered MLM. These two centering approaches were less biased only when the within-level factor loading for each item was equal to or above 0.80. Not surprisingly, because the between-level variance was not partitioned out in grand-mean centered MLM, the boxplots in each level of the within-level factor loadings across different ICC_X roughly indicated that the higher the ICC_X, the higher the relative parameter bias.

Figure 6 shows the evaluation results of the relative parameter bias across conditions for each centering approach by the level of within-level and between-level factor loadings. The overall patterns in Figure 6 are similar to the results of Figure 5. Group-mean centering and latent-mean centering are recommended only when the within-level factor loadings are equal to or above 0.80. Again, for grand-mean centering, the between-level variance cannot be separated out. The higher level of factor loadings at the within level and between level lead to an inflated VAR_M. As a result, an inflated VAR_M brings about an overestimated β_W. Therefore, in Figure 6, grand-mean centered MLM shows high relative parameter biases when the factor loadings at the between- and within-levels are also high.

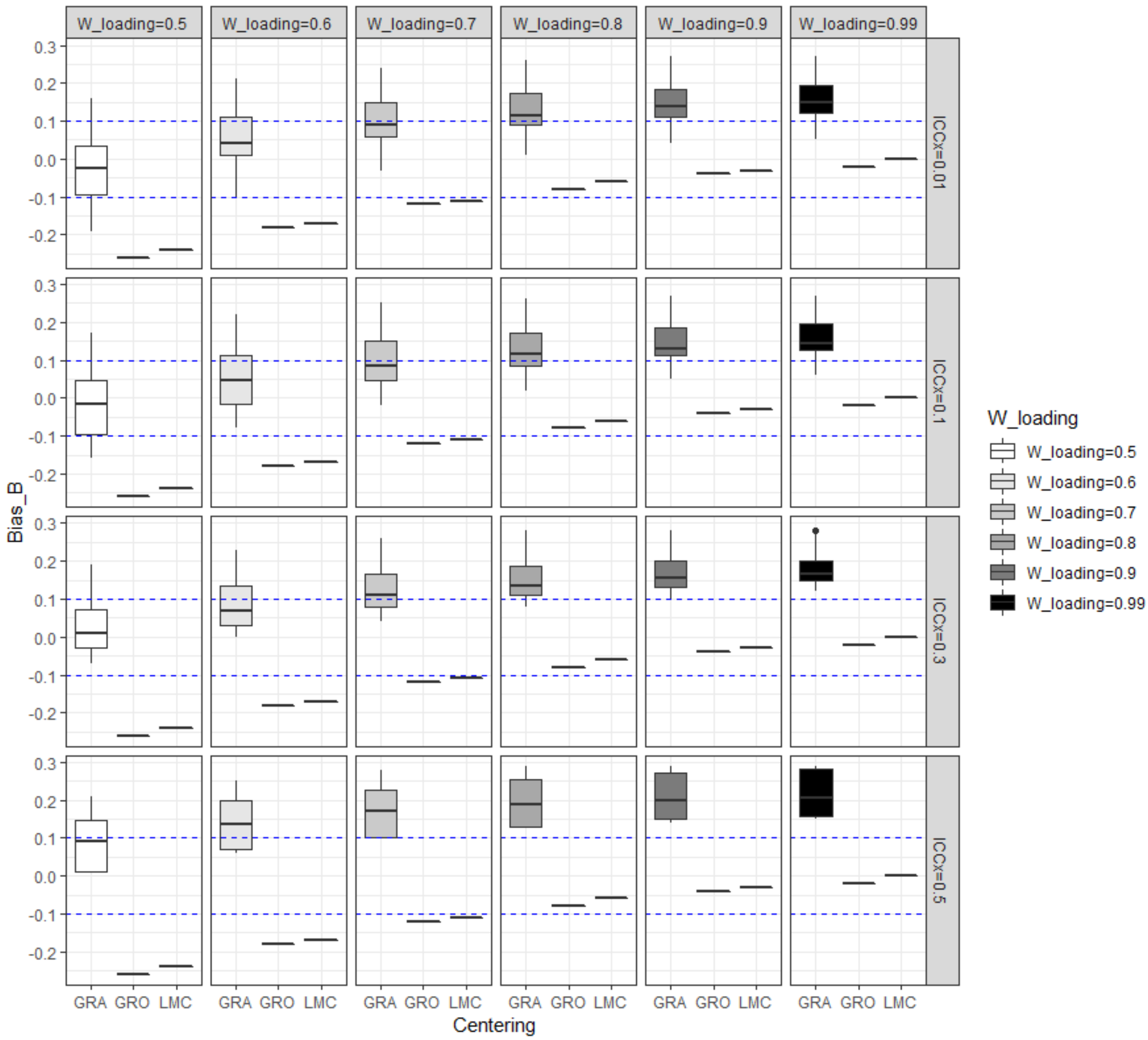

Figure 5. Boxplots showing the distribution of relative parameter bias values across conditions for each centering strategy in MLM by the level of within-level factor loadings and ICC_X.

Note. Bias_B = relative parameter bias of β_W; W_loading = within-level factor loading; GRA = grand-mean centering; GRO = group-mean centering; LMC = latent-mean centering.

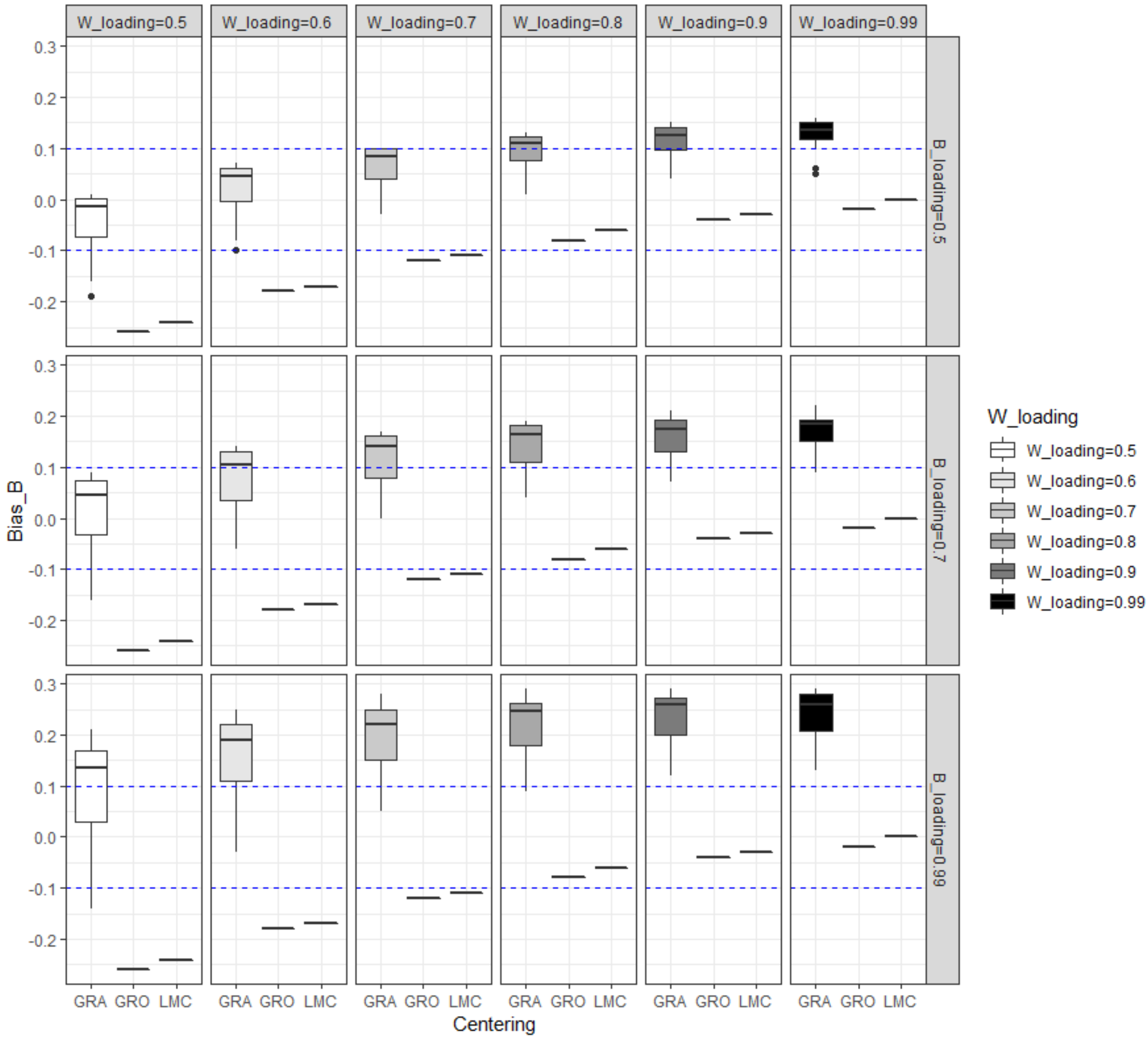

Figure 6. Boxplots showing the distribution of relative parameter bias values across conditions for each centering strategy in MLM by the level of within-level factor loadings and between-level factor loadings.

Note. Bias_B = relative parameter bias of β_W; W_loading = within-level factor loading; B_loading = between-level factor loading; GRA = grand-mean centering; GRO = group-mean centering; LMC = latent-mean centering.

Relative Standard Error Bias

The boxplots in Figure 7 illustrate the distribution of relative standard error bias values across conditions for each centering strategy in MLM by the level of within-level factor loadings. A range between the two blue dashed lines (i.e., between -0.1 and 0.1) indicates an acceptable level. Overall, when the within-level factor loading for each item was equal to or above 0.80, the relative standard error bias values for the group-mean centering and latent-mean centering across conditions were acceptable. However, most conditions under the grand-mean centering were out of the acceptable range.

Next, I examined the distribution for each centering strategy by the level of within-level factor loadings and ICC_X. Similar findings are found in Figure 8. For group-mean centering and latent-mean centering, all relative standard error bias values were within the acceptable range. However, the standard error estimates for most conditions under the grand-mean centering were biased. Specifically, as the ICC_X increased, the standard error estimates tended to be more underestimated.

Figure 9 shows the evaluation results of relative standard error bias across conditions for each centering approach by the level of within-level factor loadings and between-level factor loadings. As expected, the relative standard error bias values of group-mean centering and latent-mean centering were acceptable, while most conditions under the grand-mean centering showed an underestimated standard error especially for the conditions of high between-level factor loadings (above .70).

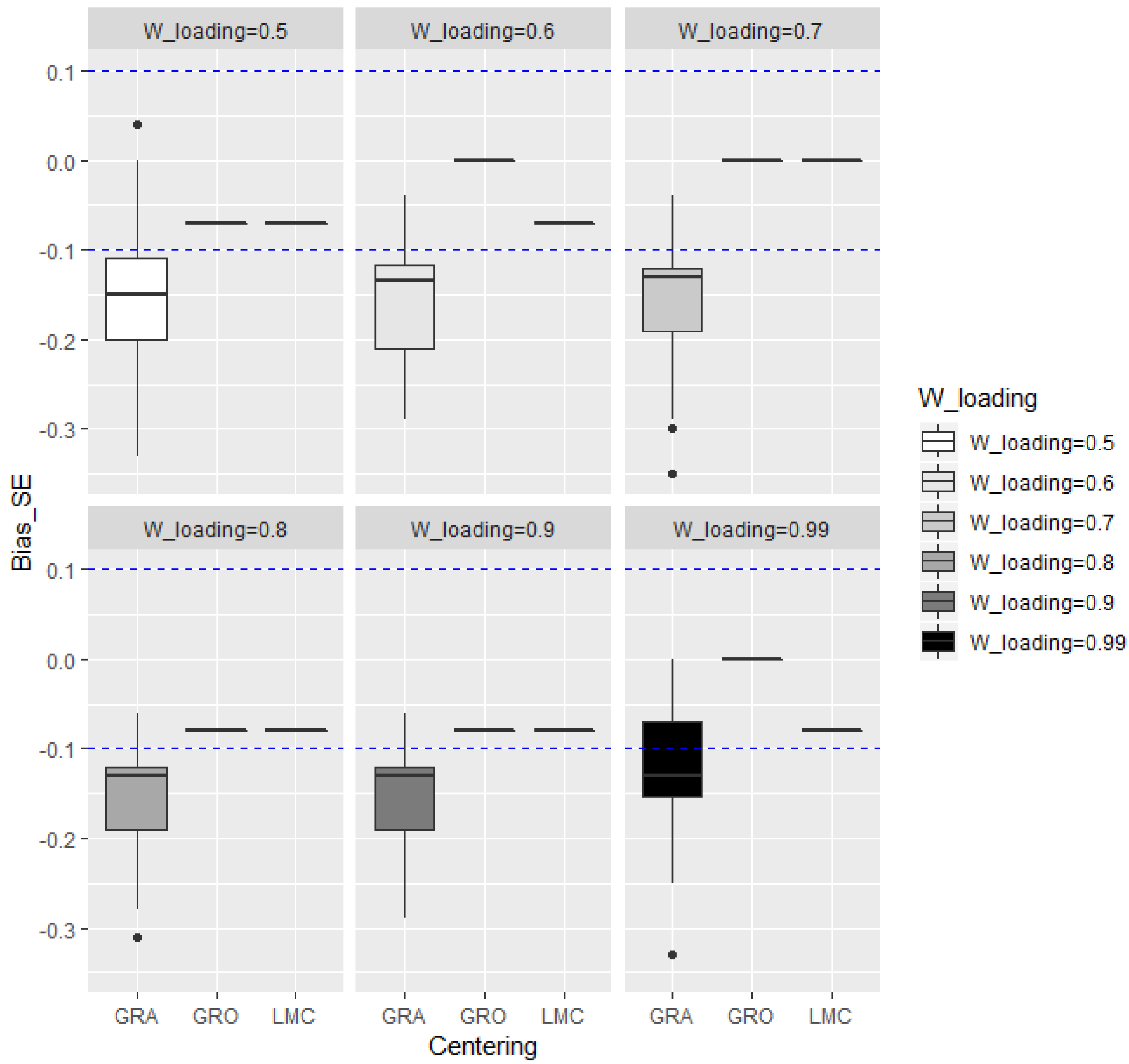

Figure 7. Boxplots showing the distribution of relative standard error bias values across conditions for each centering strategy in MLM by the level of within-level factor loadings.

Note. Bias_SE = relative standard error bias; W_loading = within-level factor loading; GRA = grand-mean centering; GRO = group-mean centering; LMC = latent-mean centering.

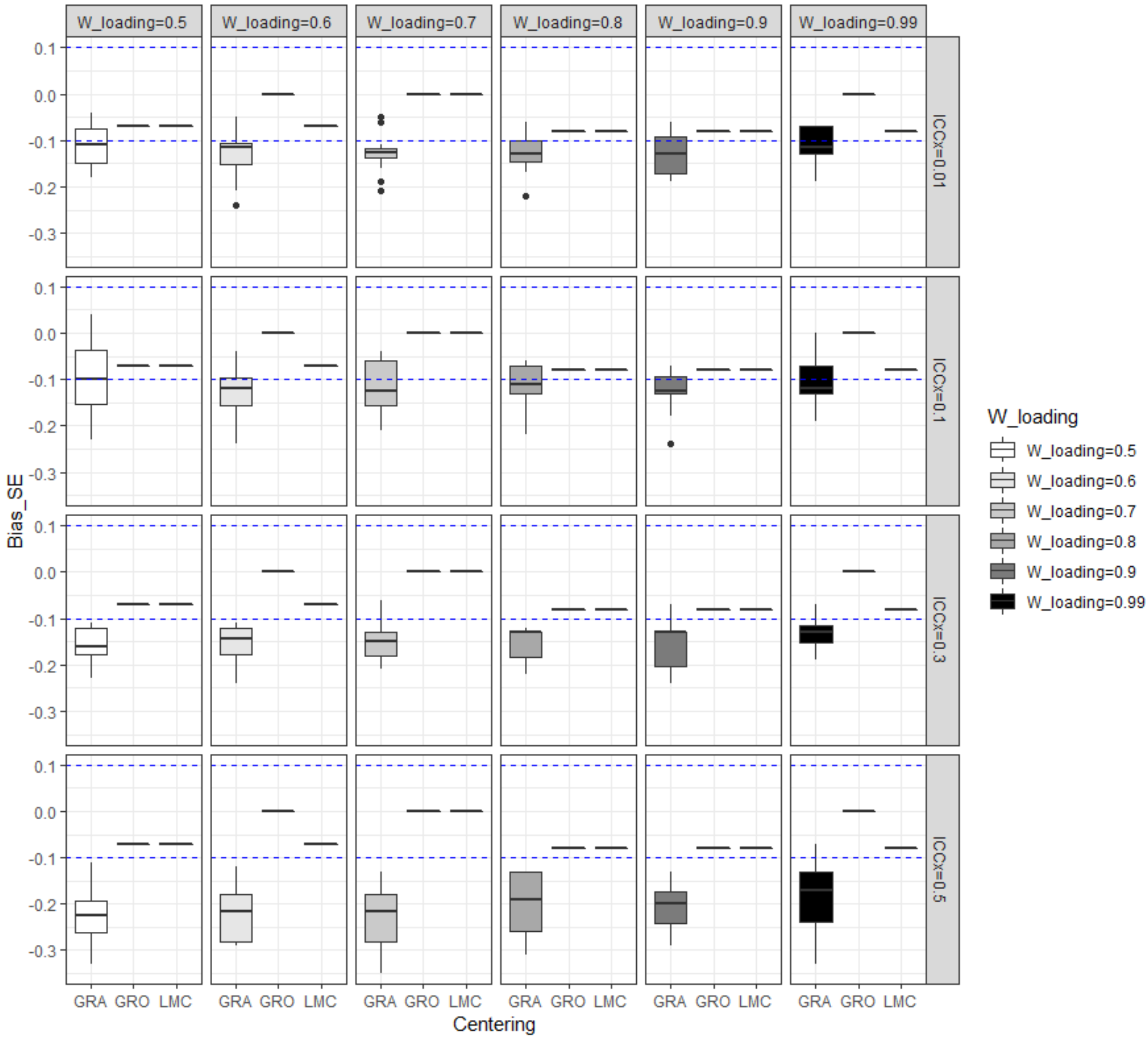

Figure 8. Boxplots showing the distribution of relative standard error bias values across conditions for each centering strategy in MLM by the level of within-level factor loadings and *ICC*$_X$.

Note. Bias_SE = relative standard error bias; W_loading = within-level factor loading; GRA = grand-mean centering; GRO = group-mean centering; LMC = latent-mean centering.

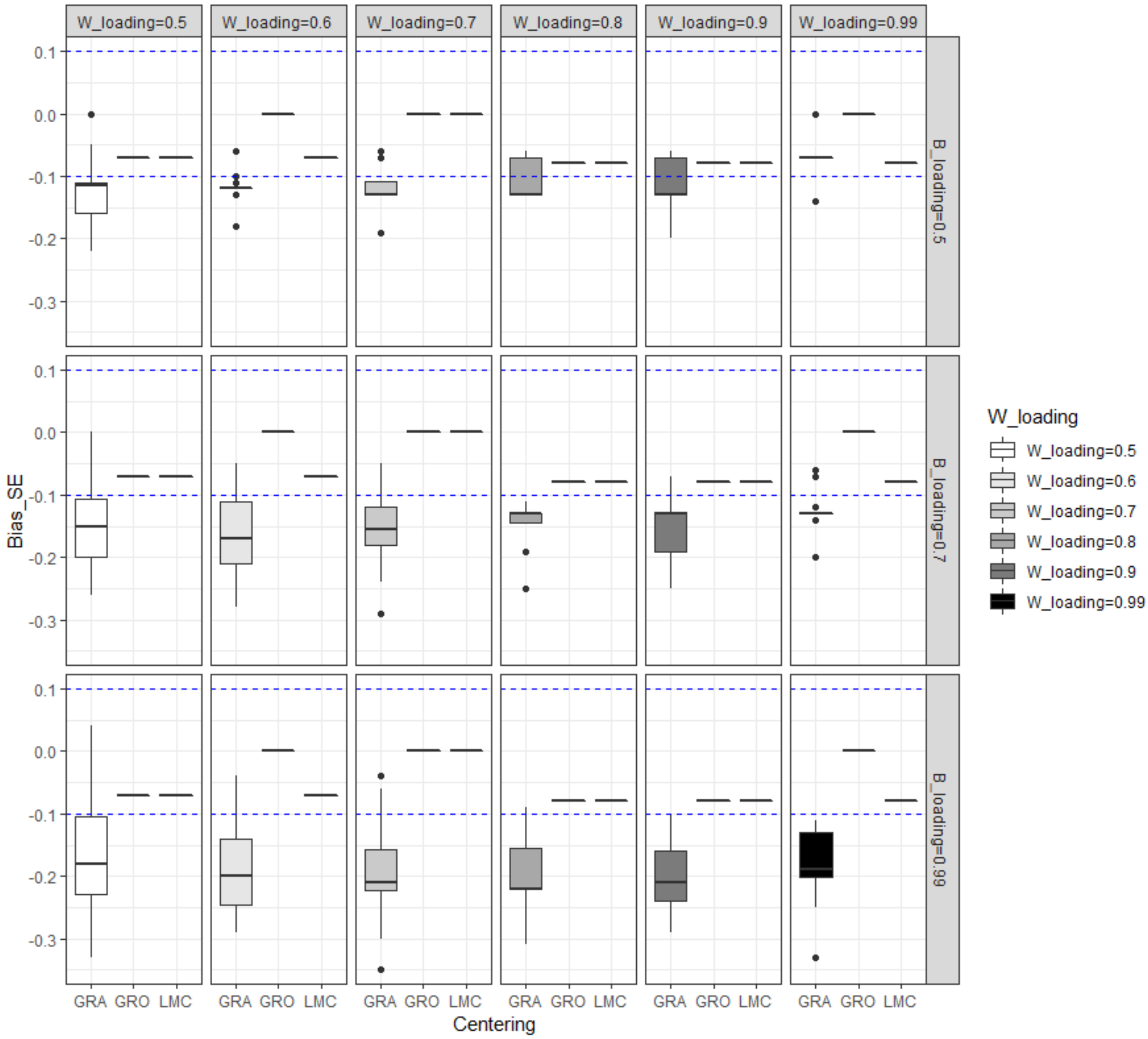

Figure 9. Boxplots showing the distribution of relative standard error bias values across conditions for each centering strategy in MLM by the level of within-level factor loadings and between-level factor loadings.

Note. Bias_SE = relative standard error bias; W_loading = within-level factor loading; B_loading = between-level factor loading; GRA = grand-mean centering; GRO = group-mean centering; LMC = latent-mean centering.

Root Mean Squared Error (*RMSE*)

The root mean square error (*RMSE*) is a measure of the overall accuracy of β_W. A larger *RMSE* indicates less accuracy in the estimate. The boxplots (Figure 10) show the distribution of *RMSE* values across conditions for each centering strategy by the level of within-level factor loadings.

As anticipated, for the group-mean centering and latent-mean centering, the higher the within-level factor loadings (i.e., the lower measurement errors), the higher the accuracy in the estimate. It was because these two centering approaches cannot deal with the measurement error issues. Overall, when the within-level factor loadings were equal to or above .80, group-mean centering and latent-mean centering showed the higher accuracy than grand-mean centering for most simulation conditions.

As for grand-mean centering, the boxplots in Figure 10 reveal the higher the within-level factor loadings (i.e., lower measurement errors), the lower the accuracy in the estimate. When the within-level factor loadings were .99 (with very few measurement errors), the grand-mean centering showed the lowest accuracy because the between-level variance showing a strong impact on the estimation was not separated out. As the within-level factor loadings decreased from .99 to .50, the decreasing within-level variance reduced the bias resulting from the between-level variance. Therefore, in some situations (e.g., within-level factor loadings = .50), grand-mean centering showed better accuracy than group-mean centering and latent-mean centering. This can be deceptive because measurement errors and data dependency could counterbalance each other at some points.

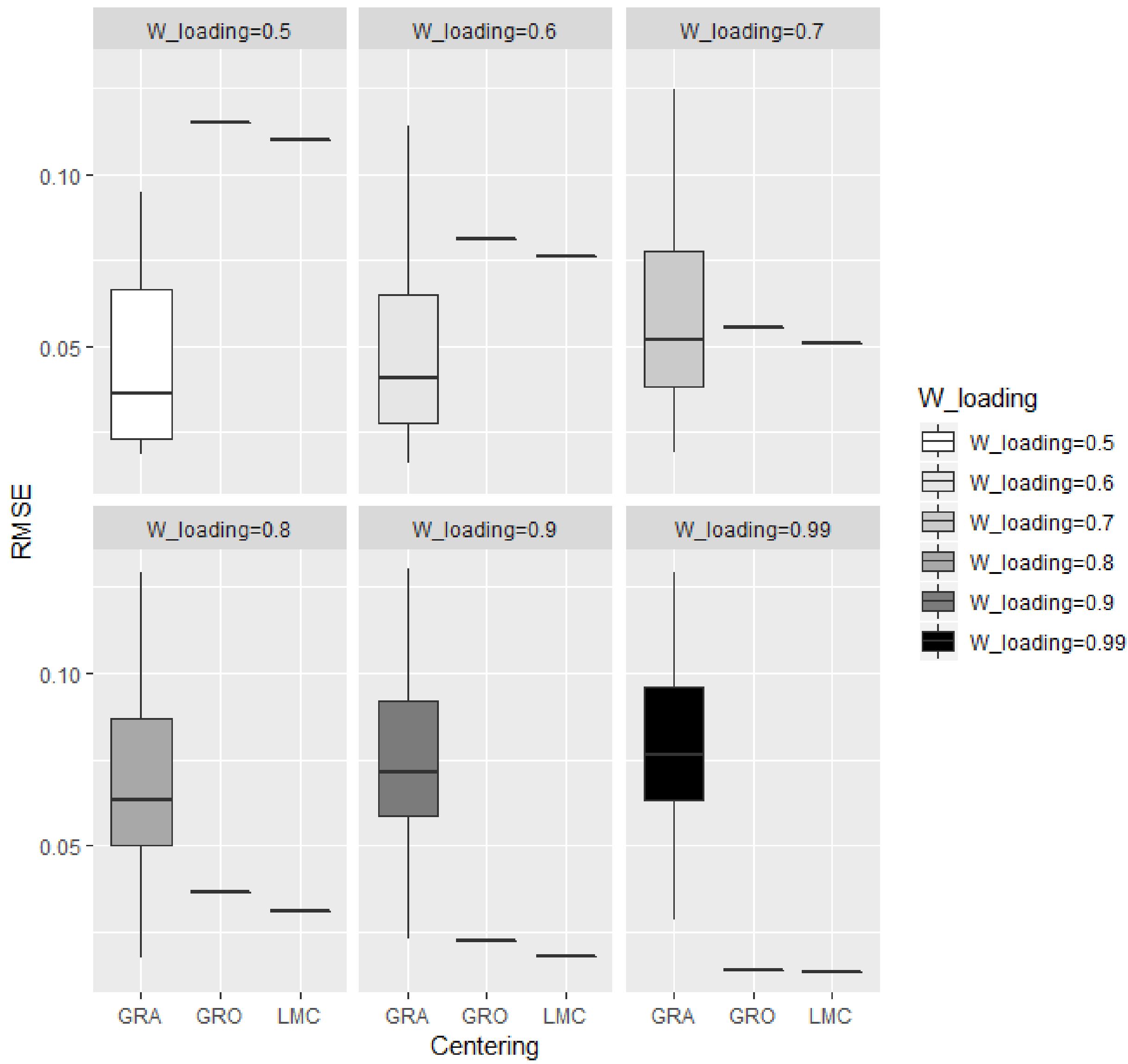

Figure 10. Boxplots showing the distribution of root mean square error (*RMSE*) values across conditions for each centering strategy in MLM by the level of within-level factor loadings.

Note. W_loading = within-level factor loading; GRA = grand-mean centering; GRO = group-mean centering; LMC = latent-mean centering.

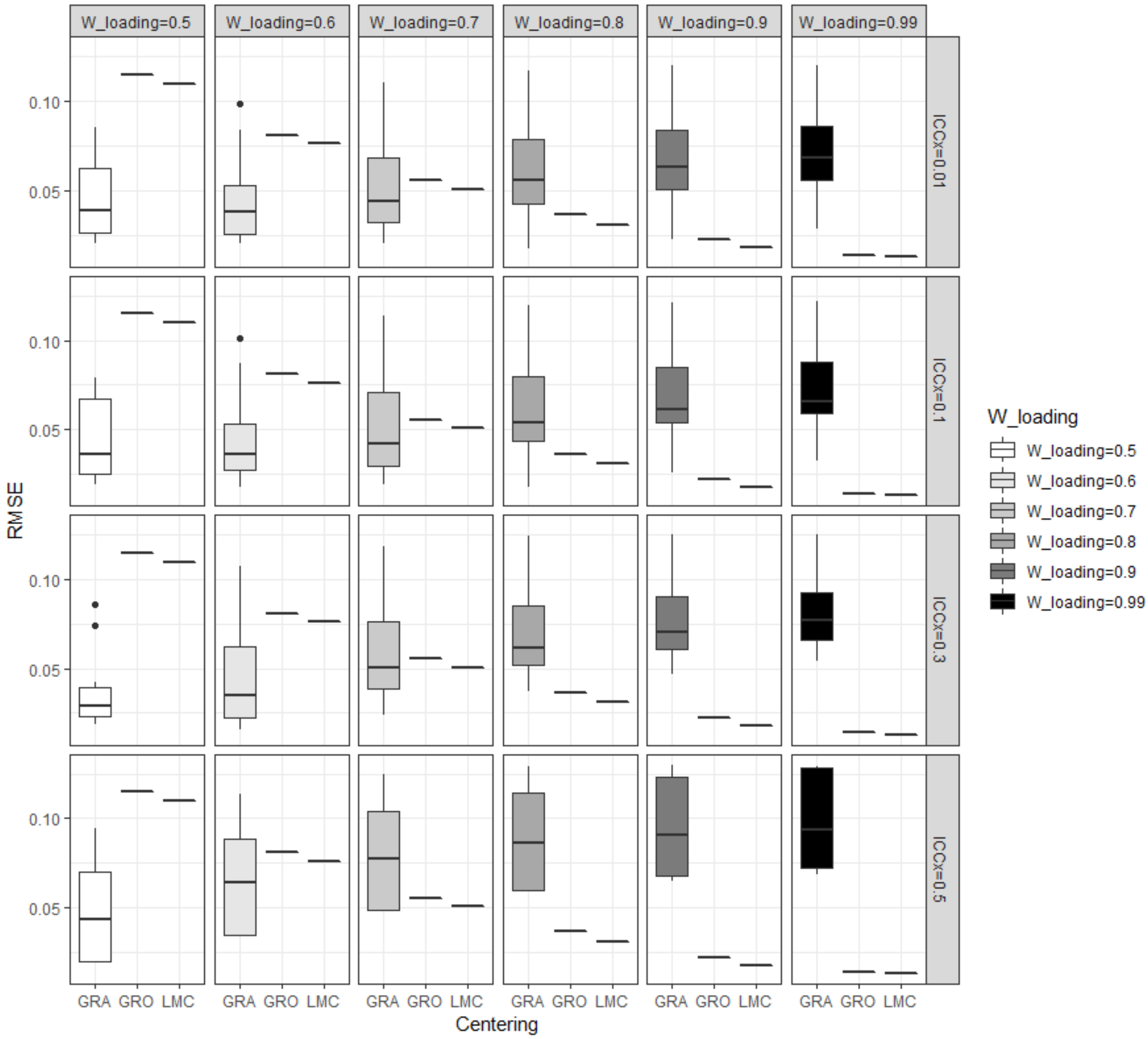

Figure 11. Boxplots showing the distribution of root mean square error (*RMSE*) values across conditions for each centering strategy in MLM by the level of within-level factor loadings and ICC_X.

Note. W_loading = within-level factor loading; GRA = grand-mean centering; GRO = group-mean centering; LMC = latent-mean centering.

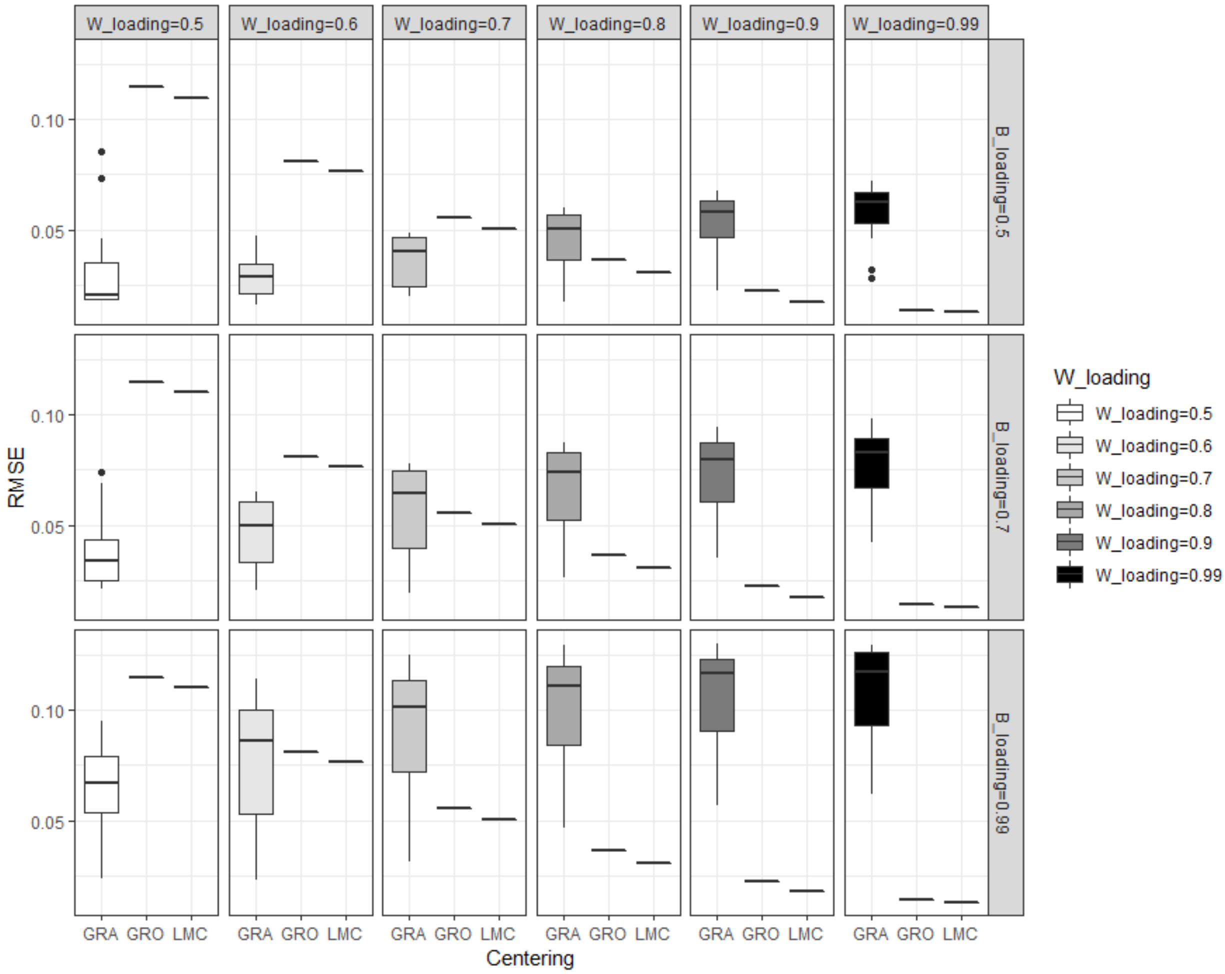

Figure 12. Boxplots showing the distribution of root mean square error (*RMSE*) values across conditions for each centering strategy in MLM by the level of within-level factor loadings and between-level factor loadings.

Note. W_loading = within-level factor loading; B_loading = between-level factor loading; GRA = grand-mean centering; GRO = group-mean centering; LMC = latent-mean centering.

Similar patterns can be found in Figures 11 and 12. Group-mean centering and latent-mean centering showed a higher accuracy than grand-mean centering for most simulation conditions when within-level factor loadings are equal to or above .80. While the within-level factor loadings were .50, grand-mean centering showed better accuracy than group-mean centering and latent-mean centering. Under the conditions of each within-level factor loading, as the ICC_X or between-level factor loadings increased, the accuracy for grand-mean centering decreased.

In sum, based on the evaluation results of the relative parameter bias, relative standard error bias, and *RMSE*, group-mean centered and latent-mean centered MLM were less biased only when the within-level factor loadings were equal to or above 0.8. Even though the grand-mean centered MLM showed better accuracy (*RMSE*) than the group-mean centered and latent-mean centered MLM in some situations (e.g., within-level factor loading = 0.5), the grand-mean centered MLM yielded unacceptable relative parameter bias and relative standard error bias under the most conditions. Accordingly, grand-mean centered MLM (or uncentered MLM) is not recommended.

CHAPTER V

DISCUSSION

Educational research is inherently multilevel. Given that students are nested within schools, students in each school sharing the same culture, resources, and experiences tend to provide researchers with similar responses. This response pattern within a school (i.e., data dependency) violates the independence assumption of ordinary least squares (OLS) regression. Therefore, MLM has become a widely used approach for analyzing multilevel data in educational research.

However, MLM still has limitations. First, MLM cannot handle the data dependency issue in the within-level predictors, and simply assumes that all the within-level predictors' ICCs are equal to zero (no data dependency issues). Second, MLM cannot include a measurement model to handle measurement errors and construct a latent factor. An alternative approach is to compute a composite score through a set of observed items from a scale, and utilize the composite score that assumes the measurement to be error free to represent it in the analysis. Computing an average score to represent a latent factor is a common approach in educational studies, especially for studies using the TIMSS data (Martin & Preuschoff, 2008; Mullis, Martin, & Foy, 2008). Although MSEM is a promising approach for dealing with these issues (Table 8), educational researchers still prefer MLM to MSEM. Therefore, this simulation study aimed to demonstrate the biased estimates that emerge from failing to account for measurement errors and data dependency in the within-level predictors in MLM.

Table 8

Modeling Latent Factors in Multilevel Settings: MLM vs. MSEM

	MLM	MSEM
Measurement Models		✓
Measurement Errors		✓
Data Dependency in Outcome	✓	✓
Data Dependency in the Within-level Predictors		✓

The regression results and Equation (11) pointed out the importance of estimating an accurate variance for the within-level latent composite predictor (VAR_M). The results indicated a higher VAR_M leads to a higher β_W. In other words, to avoid obtaining a biased β_W, VAR_M should be estimated accurately. The VAR_M consists of the within-level variance and between-level variance. The total variance of each level is computed based on factor loadings, measurement errors, and factor variances, as shown in Equation (5). The data dependency and measurement error issues potentially lead to a biased VAR_M:

(1) Data dependency

The within-level predictor in MLM is theoretically assumed to contain no between-level variance (i.e., $ICC_X = 0$). If VAR_M actually contains between-level variance, the inflated VAR_M will lead to an overestimated β_W.

(2) Measurement errors

The within-level predictor in MLM is theoretically assumed to be free of measurement error. If VAR_M contains measurement errors, which results in a smaller VAR_M, see Equation (5), the smaller VAR_M will lead to an underestimated β_W (this effect is known as regression dilution bias, or attenuation) (Hutcheon, Chiolero, & Hanley, 2010).

(3) Data dependency + Measurement errors

Data dependency implies the VAR_M contains between-level variance, leading to a large VAR_M. Measurement errors yield a small VAR_M. As discussed in Chapter II, there is a chance that data dependency (large VAR_M) and measurement errors (small VAR_M) may counterbalance each other in some situations. Therefore, using a composite score in MLM involves very complicated methodological issues, leading to an unreliable, risky result. Hence, the simulation conditions should consider the level of the intraclass correlation coefficients (ICC) for the predictor and outcome, the level of factor loadings of the latent predictor at the between- and within-levels, as well as centering strategies.

The simulation results indicated that when both data dependency and measurement error issues are involved in MLM, group-mean centering and latent-mean centering demonstrated the capacity to partition the variance into the between level and within level across various conditions. However, these two centering approaches showed the limitations of handling measurement errors. The cutoff criterion of measurement errors for group-mean centering and latent-mean centering should be each a within-level factor loading ≥ 0.8 (i.e., each measurement error variance < 0.36; within-level composite reliability $\omega \geq 0.88$) (McDonald, 1970, 1999).

In some situations (e.g., within-level factor loading = 0.5)., grand-mean centered MLM showed better accuracy (*RMSE*) than the group-mean centered and latent-mean centered options. It was because measurement errors and data dependency luckily canceled out each other. Despite that, the grand-mean centered MLM yielded unacceptable relative parameter bias and relative standard error bias under the most conditions. Hence, grand-mean centered MLM (or uncentered MLM) is not recommended.

Given that the impact from TIMSS research could potentially ripple through education policy and curriculum decisions in each country, a methodological study to evaluate the use of an average composite variable in MLM is needed. Accordingly, this simulation study followed the multilevel settings of TIMSS (i.e., the cluster size of TIMSS is about 30; the average number of clusters is approximately 150). Four items for a latent factor in this study were based on the number of items for math/science self-concept in TIMSS (Liou, 2014a, 2014b). The simulation results yielded biased β_W estimates when grand-mean centering was used. However, as reviewed in Chapter II, many recent TIMSS researchers still used the grand-mean centered MLM in their BFLPE studies, and their studies had been highly cited by other research.

There are a few limitations to this study. First, the simulation design was based on the TIMSS settings (i.e., cluster size = 30; the number of clusters = 150; the number of items = 4). Researchers in substantive areas may be situated in settings where other cluster sizes, numbers of clusters, and numbers of items would be more appropriate. Second, this study focused on simple model design, the relationship between a within-level predictor and an outcome in a two-level data structure. Researchers may have more complicated designs in their studies. Third, the current study did not address the missing data issue, and the current MLM design only focused on random intercepts with no random slopes. Therefore, the results of the group-mean centering and latent-mean centering were very similar (Asparouhov & Muthén, 2019). Future studies may consider other multilevel settings and complex models.

Overall, this study makes a number of methodological and practical contributions to the literature. For the methodological contributions, the study presented and discussed the importance of estimating an accurate variance for a within-level composite predictor when the data dependency and measurement error issues involved in the MLM analysis. For the practical contributions, the study provided researchers (especially for the TIMSS researchers) with the following two criteria of using an average composite score to represent a within-level latent predictor in MLM. First, group-mean centering or latent-mean centering must be used to deal with the data dependency of within-level predictors. Second, regarding the measurement error issue, the within-level factor loadings should be equal to or above .80 (i.e., each measurement error variance < 0.36; within-level composite reliability $\omega \geq 0.88$) (McDonald, 1970, 1999). Otherwise, MSEM is recommended.

CHAPTER VI

CONCLUSION

MSEM is a promising approach to dealing with data dependency and measurement error issues. However, many educational researchers still prefer MLM to MSEM. MLM cannot handle the data dependency issue in the within-level predictors and cannot include a measurement model to handle measurement errors and construct a latent factor. As such, computing an average score to represent a latent factor in MLM is a common alternative approach in educational studies. The current study evaluated the consequences of using an average score to represent a latent factor in MLM. The results suggested that the bias of using an average score to represent a latent predictor in MLM is acceptable only when the following criterion are met: (1) group-mean centering or latent-mean centering is utilized; (2) the within-level factor loading of each item is equal to or above .80 (i.e., within-level composite reliability $\omega \geq 0.88$). Otherwise, MSEM is recommended.

REFERENCES

Asparouhov, T., & Muthén, B. (2019). Latent variable centering of predictors and mediators in multilevel and time-series models. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(1), 119–142.

Curran, P. J. (2003). Have multilevel models been structural equation models all along? *Multivariate Behavioral Research*, *38*(4), 529–569.

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, *12*(2), 121–138.

Guo, J., Marsh, H. W., Parker, P. D., & Dicke, T. (2018). Cross-cultural generalizability of social and dimensional comparison effects on reading, math, and science self-concepts for primary school students using the combined PIRLS and TIMSS data. *Learning and Instruction*, *58*, 210–219.

Hallquist, M. N., & Wiley, J. F. (2018). M*plus*Automation: An R Package for Facilitating Large-Scale Latent Variable Analyses in M*plus*. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(4), 621–638.

Hox, J. J. (1995). *Applied multilevel analysis*. Amsterdam, The Netherlands: TT-Publikaties.

Hox, J. J. (2013). Multilevel regression and multilevel structural equation modeling. In T. D. Little (Ed.), *The Oxford handbook of quantitative methods in Psychology* (Vol. 2, pp. 281–294). Oxford, England: Oxford University Press.

Hsiao, Y.-Y., Kwok, O.-M., & Lai, M. H. (2018). Evaluation of two methods for modeling measurement errors when testing interaction effects with observed composite scores. *Educational and Psychological Measurement*, *78*(2), 181–202.

Hsu, H.-Y., Lin, J.-H., Kwok, O.-M., Acosta, S., & Willson, V. (2016). The impact of intraclass correlation on the effectiveness of level-specific fit indices in multilevel structural equation modeling: A Monte Carlo study. *Educational and Psychological Measurement*, *77*(1), 5–31.

Hutcheon, J. A., Chiolero, A., & Hanley, J. A. (2010). Random measurement error and regression dilution bias. *British Medical Journal*, *340*, 1402-1406.

IEA TIMSS & PIRLS International Study Center. (2019). Data to improve education worldwide. Retrieved August 1, 2019, from https://timssandpirls.bc.edu/index.html

Jaccard, J., & Wan, C. K. (1995). Measurement error in the analysis of interaction effects between continuous predictors using multiple regression: Multiple indicator and structural equation approaches. *Psychological Bulletin*, *117*(2), 348.

Li, X., & Beretvas, S. N. (2013). Sample size limits for estimating upper level mediation models using multilevel SEM. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*(2), 241–264.

Liou, P.-Y. (2014a). Examining the Big-Fish-Little-Pond Effect on students' self-concept of learning science in Taiwan based on the TIMSS databases. *International Journal of Science Education*, *36*(12), 2009–2028.

Liou, P.-Y. (2014b). Investigation of the big-fish-little-pond effect on students' self-concept of learning mathematics and science in Taiwan: Results from TIMSS 2011. *The Asia-Pacific Education Researcher*, *23*(3), 769–778.

Liou, P.-Y., & Jessie Ho, H.-N. (2018). Relationships among instructional practices, students' motivational beliefs and science achievement in Taiwan using hierarchical linear modelling. *Research Papers in Education*, *33*(1), 73–88.

Ma, J., Raina, P., Beyene, J., & Thabane, L. (2012). Comparing the performance of different multiple imputation strategies for missing binary outcomes in cluster randomized trials: A simulation study. *J Open Access Med Stat*, *2*, 93–103.

Marsh, H. W., Abduljabbar, A. S., Parker, P. D., Morin, A. J., Abdelfattah, F., & Nagengast, B. (2014). The Big-Fish-Little-Pond effect in mathematics: A cross-cultural comparison of US and Saudi Arabian TIMSS responses. *Journal of Cross-Cultural Psychology*, *45*(5), 777–804.

Marsh, H. W., & Parker, J. W. (1984). Determinants of student self-concept: Is it better to be a relatively large fish in a small pond even if you don't learn to swim as well? *Journal of Personality and Social Psychology*, *47*(1), 213–231.

Martin, M. O., & Preuschoff, C. (2008). Chapter 12: Creating the TIMSS 2007 background indices. In J. F. Olson, M. O. Martin, & I. V. Mullis (Ed.), *TIMSS 2007 technical report* (pp. 281–338). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

McDonald, R. P. (1970). The theoretical foundations of principal factor analysis, canonical factor analysis, and alpha factor analysis. *British Journal of Mathematical and Statistical Psychology*, *23*(1), 1-21.

McDonald, R. P., & Goldstein, H. (1989). Balanced versus unbalanced designs for linear structural relations in two-level data. *British Journal of Mathematical and Statistical Psychology*, *42*(2), 215–232.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.

Min, I., Cortina, K. S., & Miller, K. F. (2016). Modesty bias and the attitude-achievement paradox across nations: A reanalysis of TIMSS. *Learning and Individual Differences*, *51*, 359–366.

Mohammadpour, E. (2012a). A multilevel study on trends in Malaysian secondary school students' science achievement and associated school and student predictors. *Science Education*, *96*(6), 1013–1046.

Mohammadpour, E. (2012b). Factors accounting for mathematics achievement of Singaporean eighth-graders. *The Asia-Pacific Education Researcher*, *21*(3), 507–518.

Mohammadpour, E. (2013). A three-level multilevel analysis of Singaporean eighth-graders science achievement. *Learning and Individual Differences, 26*, 212–220.

Mohammadpour, E., & Abdul Ghafar, M. N. (2014). Mathematics achievement as a function of within-and between-school differences. *Scandinavian Journal of Educational Research*, *58*(2), 189–221.

Mullis, I. V. S., Martin, M. O., & Foy, P. (2008). Chapter 4: Students' backgrounds and attitudes towards mathematics. *TIMSS 2007 international mathematics report* (pp. 143-188). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Muthén, B. (1989). Multiple-group structural modelling with non-normal continuous variables. *British Journal of Mathematical and Statistical Psychology*, *42*(1), 55–62.

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research*, *22*(3), 376–398.

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus User's Guide* (Eighth Edition). Los Angeles, CA: Muthén & Muthén.

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

Preacher, K. J., Zhang, Z., & Zyphur, M. J. (2011). Alternative methods for assessing mediation in multilevel data: The advantages of multilevel SEM. *Structural Equation Modeling*, *18*(2), 161–182.

Preacher, K. J., Zyphur, M. J., & Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods, 15*(3), 209.

Raudenbush, S., & Bryk, A. S. (1986). A hierarchical model for studying school effects. *Sociology of Education*, *59*(1), 1–17.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Second Edition). Newbury Park, CA: Sage.

Rose, N., Wagner, W., Mayer, A., & Nagengast, B. (2019). Model-Based Manifest and Latent Composite Scores in Structural Equation Models. *Collabra: Psychology*, *5*(1).

Ryu, E. (2015). The role of centering for interaction of level 1 variables in multilevel structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*(4), 617–630.

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

Schneider, B., Krajcik, J., Lavonen, J., Salmela-Aro, K., Broda, M., Spicer, J., … Juuti, K. (2016). Investigating optimal learning moments in US and Finnish science classes. *Journal of Research in Science Teaching*, *53*(3), 400–421.

Sellström, E., & Bremberg, S. (2006). Is there a “school effect” on pupil outcomes? A review of multilevel studies. *Journal of Epidemiology & Community Health*, *60*(2), 149-155.

Sheldrake, R. (2016). Differential predictors of under-confidence and over-confidence for mathematics and science students in England. *Learning and Individual Differences*, *49*, 305–313.

Shin, Y., & Raudenbush, S. W. (2010). A latent cluster-mean approach to the contextual effects model with missing data. *Journal of Educational and Behavioral Statistics*, *35*(1), 26–53.

Sullivan, L. M., Dukes, K. A., & Losina, E. (1999). An introduction to hierarchical linear modelling. *Statistics in Medicine*, *18*(7), 855–888.

Tsai, L.-T., & Yang, C.-C. (2015). Hierarchical effects of school-, classroom-, and student-level factors on the science performance of eighth-grade Taiwanese students. *International Journal of Science Education*, *37*(8), 1166–1181.

Wang, C.-L., & Liou, P.-Y. (2017). Students' motivational beliefs in science learning, school motivational contexts, and science achievement in Taiwan. *International Journal of Science Education*, *39*(7), 898–917.

Wang, X. (2013). Why students choose STEM majors: Motivation, high school learning, and postsecondary context of support. *American Educational Research Journal*, *50*(5), 1081–1121.

Wang, Z. (2015). Examining big-fish-little-pond-effects across 49 countries: A multilevel latent variable modelling approach. *Educational Psychology*, *35*(2), 228–251.

Wang, Z., & Bergin, D. A. (2017). Perceived relative standing and the big-fish-little-pond effect in 59 countries and regions: Analysis of TIMSS 2011 data. *Learning and Individual Differences*, *57*, 141–156.

Wilkins, J. L. (2004). Mathematics and science self-concept: An international investigation. *The Journal of Experimental Education*, *72*(4), 331–346.

Woltman, H., Feldstain, A., MacKay, J. C., & Rocchi, M. (2012). An introduction to hierarchical linear modeling. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 52–69.