ABSTRACT

Title of dissertation: A FINITE MIXTURE MULTILEVEL STRUCTURAL EQUATION MODEL FOR UNOBSERVED HETEROGENEITY IN RANDOM VARIABILITY

Yi Feng, Doctor of Philosophy, 2023

Dissertation directed by: Professor Gregory R. Hancock
Program in Quantitative Methodology: Measurement and Statistics
Department of Human Development and Quantitative Methodology

Variability is often of key interest in various research and applied settings. Important research questions about *intraindividual variability* (e.g., consistency across repeated measurements) or *intragroup variability* (e.g., cohesiveness among members within a team) are piquing the interest of researchers from a variety of disciplines. To address the research needs in modeling random variability as the key construct, Feng and Hancock (2020, 2022) proposed a multilevel SEM-based modeling approach where variability can be modeled as a random variable. This modeling framework is a highly flexible analytical tool that can model variability in observed measures or latent constructs, variability as the predictor or the outcome, as well as the between-subject comparison of variability across observed groups. A huge challenge still remains, however, when it comes to modeling the unobserved heterogeneity in random variability. Given that no existing research addresses the methodological considerations of uncovering the unobserved sub-populations that differ in *intraindividual variability* or

intragroup variability, or sub-populations that differ in the various processes and mechanisms involving *intraindividual variability* or *intragroup variability*, the current dissertation study aims to fill this gap in literature. In the current study, a finite-mixture MSEM for modeling unobserved heterogeneity in random variability (MMSEM-RV) is introduced. Bayesian estimation via MCMC is proposed for model estimation. The performance of MMSEM-RV with Bayesian estimation is systematically evaluated in a simulation study across varying conditions. An illustrative example with empirical PISA data is also provided to demonstrate the practical application of MMSEM-RV.

A FINITE MIXTURE MULTILEVEL STRUCTURAL EQUATION MODEL FOR UNOBSERVED HETEROGENEITY IN RANDOM VARIABILITY

by

Yi Feng

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2023

Advisory Committee:

Professor Gregory R. Hancock, Chair & Advisor
Professor Jeffrey R. Harring
Professor Laura M. Stapleton
Professor Peter M. Steiner
Professor Tracy Riggins

© Copyright by
Yi Feng
2023

Dedication

To my beloved grandparents, who have been there guiding me since my very first steps, the first time I wrote with a pen, and the first math problem I ever solved. My education would not be complete without your support and unconditional love.

To my dearest academic family at QMMS, my sanctuary, my second home, and the extraordinary souls in it who have become my Ohana along this journey. I have been able to grow into a better scholar and a better person being part of this family.

Acknowledgements

I have been incredibly fortunate to have received extensive help and support throughout the course of this dissertation study. However, my acknowledgements extend far beyond this research project. I have so many people to thank for, who have contributed profoundly to my graduate training and to shaping my growth trajectory as a quantitative methodologist.

My deepest appreciation goes to my advisor, Dr. Gregory R. Hancock, who has been a great mentor, a role model, a source of inspiration, a close friend, a dear family, and a crime partner throughout my years in graduate school. He exemplifies the best side of academia and is the best mentor one could ever ask for. During the years, while we were working together to tackle the many methodological challenges, he generously shared his wealth of wisdom and patiently guided me through any obstacles. It was through these many insightful conversations and shared fun moments that my passion and enthusiasm for advanced structural equation modeling, power analysis, and quantitative methods got inspired tremendously. Having Gregory as my advisor is knowing that I always have someone in my corner. There were times when, he would still make time to be on a call with me to provide immediate help, even at the oddest hour of his day while traveling half a world away. He has contributed to making me a better version of myself in so many ways and his positive influence certainly goes far beyond this dissertation work. I would not have been able to become who I am or reach where I am at today without his unwavering support. As I embark on the next chapter of my journey, I know his remarkable mentorship will continue to light my way ahead. I will always carry with me his invaluable advice, along with our shared dedication to the field of quantitative methodology—which is

constantly reminded by the serious voice memo his podcast intern left on my phone that later became my ringtone— “I solemnly swear that I am up to no good”.

I reserve a very special thank-you for Dr. Jeffrey Harring, whose presence throughout my time in QMMS (EDMS) has been nothing short of transformative. Jeff is like a loving parent as well as a dedicated coach, who cares deeply about student growth and invests wholeheartedly in their development. He is willing to spend hours sitting down with me, carefully going through my code line by line, and offering critical insights and valuable suggestions. It was with his exceptional expertise and patient guidance that I was able to improve my simulation and coding skills, which has become one of the cornerstones of my research in quantitative methodology. His extensive knowledge on mixture models has also been of incredible help for this dissertation study. Jeff extends his mentorship well beyond the classroom, marking my graduate journey with many sparkling memories. From the subway adventure to attending the APS conference in DC, the fun road trip to Penn State University for the longitudinal data analysis conference, to the successful organization of SMEP conference in Baltimore, he captained me through so many major academic events. Over the years, Jeff has also become a very dear friend, who is a ray of sunshine even on the cloudiest days. His infectious smile instantly brightens up the room whenever he walks in. He would drive extra miles, quite literally, to share his homemade Christmas cookies with me. His warmth and support are the glue that holds our program closely together. His kindness, wisdom, and positivity have shaped not only my graduate experience, but also my outlook on life. He is, and always will be, a true inspiration.

I owe a debt of gratitude that words can hardly express to Dr. Laura Stapleton—our queen, chair, and my academic sibling, Laura. It was Laura who first introduced me to this program, forever altering the course of my life. She was the first faculty member who believed in

my potential as a quantitative methodologist, even when I had doubts about myself. I still vividly recall the conversation in our first meeting on my recruitment day. Sensing my uncertainty, she reassured me, "Yi, but you already sound like one of us!". For as long as I am part of QMMS (EDMS), Laura always generously provided me with continuing assistantship opportunities. These assistantships not only provided the much-needed financial support for me, but, most importantly, also allowed me to work closely with her on research and learn how to connect our academic endeavors with the practical needs of local communities. To better help me prepare for my future academic career, she entrusted me with opportunities to contribute to grant proposals and meticulously guided me through the process. She also included me for co-teaching in her workshops. Despite her busy schedule, Laura has always made herself available to offer invaluable suggestions on my research work, from reviewing my first R package to helping refine this dissertation work. After my proposal defense, she handed me a copy filled with her handwritten notes and detailed suggestions, which greatly helped enhance the quality of this dissertation work. I am also grateful for her genuine insights, advice, and encouragement when it comes to the challenging decisions about my career and professional development. She is an excellent leader in every aspect of her role and will always be my role model as I move forward.

I must acknowledge Dr. Peter Steiner, who might already raise an eyebrow at me addressing him this way. Peter has important influences on my academic journey and I owe him a load of thanks. He introduced me to graphical causal models, opening up a whole new world of knowledge and possibilities for my research. While I am not his formal student, he is always generous in sharing his time, resources, expertise, and wisdom with me. Our collaborative work first started from a simple question I had while taking his class, which he took seriously and gave a lot of thoughts of. With his guidance, it then grew into conference presentations, grant

proposal, and ongoing academic publications. Our conversations are always joyful and have been nothing short of enlightening, and I have gained invaluable insights from him in this process over the past years. Peter sets an extraordinary model for working with students. He treats us all as equal peers, friends, and partners without emphasizing any hierarchical roles, fostering an environment of open communication. I have greatly benefited from this level of inclusiveness and equality, and it is something I hope to replicate when I interact with students in the future. Attending conferences along with Peter is also a huge pleasure. He is always there actively guiding us in navigating the event, and serving as a central node that connects and expands our network. Beyond all these, I also would like to extend special thanks to his private tea house, which has generously sponsored many valuable discussions over the years. Much like his brewed tea, Peter's impact is warm and enduring; his wisdom and guidance enrich those fortunate enough to learn from him.

I also would like to express my sincere gratitude to Dr. Tracy Riggins for graciously agreeing to be a part of my dissertation committee and for providing valuable insights into this study. I truly appreciate her willingness to contribute to this important milestone in my academic journey.

Table of Contents

Dedication ii

Acknowledgements iii

Table of Contents vii

List of Tables ix

List of Figures x

List of Abbreviations xii

Chapter 1. Introduction 1

1.1 Statement of the Problem 2
1.2 Purpose of the Study 7
1.3 Significance of the Study 9
1.4 Overview of the Chapters 10

Chapter 2. Literature Review 12

2.1 Modeling Random Variability 12
2.1.1 Multi-step Approach 12
2.1.2 Mixed Effects Modeling Approaches 13
2.1.3 Multilevel Structural Equation Modeling Approaches 20
2.2 Model-based Approaches for Unobserved Population Heterogeneity 29
2.2.1 Finite Mixture Models 29
2.2.2 Multilevel Mixture Models 30
2.2.3 Finite Mixture Multilevel Structural Equation Modeling 32

Chapter 3. Theoretical Framework 36

3.1 Finite Mixture MSEM for Random Variability 36
3.2 Model Estimation 41
3.2.1 Bayesian Estimation 43
3.2.2 Label Switching 46
3.2.3 Model Fit Assessment 49
3.3 Measurement Invariance 50

3.4 Model Assumptions 53

Chapter 4. Simulation Studies 55

4.1 Simulation Design 55

4.1.1 Data Generating Model 56

4.1.2 Manipulated Factors 58

4.1.3 Analytical Model and Estimation 62

4.2 Simulation Study Part I: Population Analysis 65

4.2.1 Outcome Evaluation 67

4.2.2 Results 68

4.3 Simulation Study Part II: Monte Carlo Simulation 91

4.3.1 Outcome Evaluation 91

4.3.2 Results 93

4.4 Summary of Simulation Study Results 106

Chapter 5. Applied Illustrative Example 109

5.1 Research Context and Analytical Data 109

5.2 Methods 111

5.3 Results 114

Chapter 6. Discussion 119

6.1 Conclusion 119

6.2 Limitations and Future Directions 128

Appendix A. Common Measures of Intraindividual Variability 133

Appendix B. Simulation Study Results 134

Appendix C. BUGS Code for Fitting the Latent Construct MMSEM-RV 166

Appendix D. Estimation Results of the Naive Mean Score MMSEM-RV for the Applied Illustrative Example 169

References 170

List of Tables

Table 4-1 Manipulated Factors and Levels in Simulation Study 59
Table 4-2 The Value of Composite Reliability and ICC under Different Conditions 61
Table 4-3 Monte Carlo Variance and RMSE Obtained from the Monte Carlo Simulation..... 99
Table 5-1 Parameter Priors Utilized in the Applied Illustrative Example 112
Table 5-2 Estimation Results for the Illustrative Example 117

List of Figures

Figure 1-1 Hypothetical Examples for Randomly Varying Intragroup Variability and Intraindividual Variability 4

Figure 1-2 Hypothetical Examples Where There Are Unobserved Sub-populations That Differ in Randomly Varying Intragroup Variability and Intraindividual Variability..... 6

Figure 2-1 A Path Diagram for a Two-level MSEM Model..... 22

Figure 2-2 MSEM-RV with One Within-Level Outcome Variable and One Between-Level Latent Outcome Variable..... 28

Figure 2-3 Path Diagram for Between-Class Multilevel Factor Mixture Model..... 34

Figure 3-1 Path Diagram for MMSEM-RV with One Common Latent Factor at the Within-Level and the Corresponding Configural Construct at the Between-Level..... 40

Figure 4-1 Distribution of the Within-Level Variance in the Simulation Study 58

Figure 4-2 Path Diagram for Unconditional Mean Score MMSEM-RV..... 65

Figure 4-3 Relative Bias for Estimating the Probability of Belonging to Class 1 71

Figure 4-4 Relative Bias for Estimating Average Within-Level Log Variance in Class 1 and Class 2..... 72

Figure 4-5 Relative Bias for Estimating the Between-Level Latent Construct Variance in Class 1 and Class 2 73

Figure 4-6 Relative Bias for Estimating the Between-Level Random Effects Covariance in Class 1 and Class 2 74

Figure 4-7 Relative Bias for Estimating Variance of the Within-Level Log Variance in Class 1 and Class 2 75

Figure 4-8 95% HPDI Coverage for Estimating the Probability of Belonging to Class 1 77

Figure 4-9 95% HPDI Coverage for Estimating Average Within-Level Log Variance in Class 1 and Class 2 78

Figure 4-10 95% HPDI Coverage for Estimating the Between-Level Latent Construct Variance in Class 1 and Class 2 79

Figure 4-11 95% HPDI Coverage for Estimating the Between-Level Random Effects Covariance in Class 1 and Class 2 80

Figure 4-12 95% HPDI Coverage for Estimating Variance of the Within-Level Log Variance in Class 1 and Class 2 81

Figure 4-13 Relative Bias for Estimating the Probability of Belonging to Class 1 under the Two Different Modeling Approaches 86

Figure 4-14 Relative Bias for Estimating Average Within-Level Log Variance in Class 1 and Class 2 under the Two Different Modeling Approaches 87

Figure 4-15 Relative Bias for Estimating the Between-Level Latent Construct Variance in Class 1 and Class 2 under the Two Different Modeling Approaches 88

Figure 4-16 Relative Bias for Estimating the Between-Level Random Effects Covariance in Class 1 and Class 2 under the Two Different Modeling Approaches..... 89

Figure 4-17 Relative Bias for Estimating Variance of the Within-Level Log Variance in Class 1 and Class 2 under the Two Different Modeling Approaches 90

Figure 4-18 Average Point Estimates with Standard Error under Condition A..... 96

Figure 4-19 Relative Bias for Focal Parameters under Condition A 97

Figure 4-20 95% HPDI Coverage Probability for Focal Parameters under Condition A..... 98

Figure 4-21 Average Point Estimates with Standard Error under Condition B..... 103

Figure 4-22 Relative Bias for Focal Parameters under Condition B 104

Figure 4-23 95% HPDI Coverage Probability for Focal Parameters under Condition B..... 105

Figure 5-1 Distribution of the Three Items Measuring Perceived Math Teacher Support within Each School for a Random Sample of 11 Schools 110

Figure 5-2 Latent Class Assignment of a Random Sample of 11 Schools 118

Figure 6-1 Scatter Plots for a Simulated Data with Small Intercept Separation and Low Factor Loading 124

Figure 6-2 Violin Plots for a Random Sample of 10 Schools in a Simulated Data with Small Intercept Separation and Low Factor Loading..... 125

Figure 6-3 Scatter Plots for a Simulated Data with Small Intercept Separation and High Factor Loading 126

Figure 6-4 Violin Plots for a Random Sample of 10 Schools in a Simulated Data with Small Intercept Separation and High Factor Loading..... 127

List of Abbreviations

AIC	Akaike information criterion
BIC	Bayesian information criterion
CFA	Confirmatory factor analysis
CFLVM	Confirmatory factor level and variability model
CV	Coefficient of variation
DHGLM	Double hierarchical generalized linear models
DSEM	Dynamic structural equation modeling
EMA	Ecological momentary assessment
FMM	Finite mixture modeling
HMC	Hamiltonian Monte Carlo
HPDI	Highest posterior density intervals
ILD	Intensive longitudinal data
iMSSD	Intraindividual mean square of successive differences
iSD	Intraindividual standard deviation
iSDr	Intraindividual residual standard deviation
MCMC	Markov Chain Monte Carlo
MELSM	Mixed effects location scale model
ML	Maximum likelihood
MLM	Multilevel modeling
MMSEM-RV	Mixture multilevel structural equation modeling for random variability
MSEM	Multilevel structural equation modeling
MMSEM-RV	Multilevel structural equation modeling for random variability
OECD	Organisation for Economic Co-operation and Development
PISA	Programme for International Student Assessment
PSR	Potential scale reduction
RMSE	Root mean squared error
SEM	Structural equation modeling
TVEM	Time-varying effects modeling

Chapter 1. Introduction

In many research and applied settings across the social, behavioral, and health sciences, it is variability, rather than averages, that is of key interest. Important research questions about *intraindividual variability* (e.g., consistency across repeated measurements) or *intragroup variability* (e.g., cohesiveness among members within a team) are drawing the interest of researchers from a variety of disciplines. To address the broad research needs in modeling random variability as the key construct, Feng and Hancock (2020, 2022) proposed a multilevel SEM-based modeling (MSEM) approach in which variability can be modeled as a random variable (MSEM-RV). This modeling framework is a highly flexible analytical tool that can model variability in observed measures or latent constructs, variability as the predictor or the outcome, as well as the between-subject comparison of variability across observed groups, thereby opening up many research possibilities that involve inquiries about variability.

In certain cases, however, the heterogeneity in intraindividual or intragroup variability is driven by the differences among *unobserved* subgroups of the population. Despite all the methodological advancements in variability modeling in recent years, there is no existing analytical tool that can be employed to identify unobserved sub-populations with varying patterns of random variability. The proposed study thus aims to address this methodological challenge by conceptually formulating and systematically evaluating the performance of a MSEM-based finite mixture model to detect the unobserved heterogeneity in the distribution of random variability.

1.1 Statement of the Problem

What is random variability and why does it matter? Variability is often of key interest in many research and applied settings. Examples of variability include consistency/stability of an individual over multiple measurements (*intraindividual variability*), and cohesiveness among members within a group or team (*intragroup variability*).

In terms of intraindividual variability, for instance, an infant often has daily fluctuations in crying behaviors and motor movement, an adult usually experiences day-to-day variations in bedtime and emotional status, and a senior tends to have varying reaction times across trials on cognitive tasks. Instead of treating such within-person fluctuations as random errors, modern developmental science theorizes that intraindividual variability contains valuable information about human development (Hamaker, 2012; van Dijk & van Geert, 2015). Importantly, it has long been acknowledged that individuals can differ in their behavioral stability or consistency (e.g., Nesselroade, 1991; Nesselroade & Ram, 2004). Because intraindividual consistency has important implications across the life span, it thus becomes an important research topic across a variety of domains, such as motor skill development, socio-emotional development, cognitive and language development, and cognitive aging (e.g. Adolph et al., 2015; Fagot et al., 2018; Hultsch & MacDonald, 2004; Kupers et al., 2019).

Methodologically parallel, intragroup variability is also of wide interest across several disciplines. Just as individuals can be different in intraindividual consistency, groups can also have varying levels of within-group dispersion, which may be further associated with important group-level outcomes. Peer heterogeneity within classrooms or schools, for example, is believed to impact student achievement (e.g., Decristan et al., 2017; Raitano & Vona, 2013), while workforce diversity more broadly (e.g., task-related attitudinal variation) and team member

cohesion are believed to have impacts on group performance and productivity (e.g., Cox & Blake, 1991; Easley, 2001; Jehn et al., 1999; Petersen et al., 2004; van Knippenberg et al., 2004). Relatedly, in biological science, intra-species variation in genotype and phenotype helps explain the difference in collective behaviors of group-living animals (e.g., Kaufhold & van Leeuwen, 2019; Knebel et al., 2019).

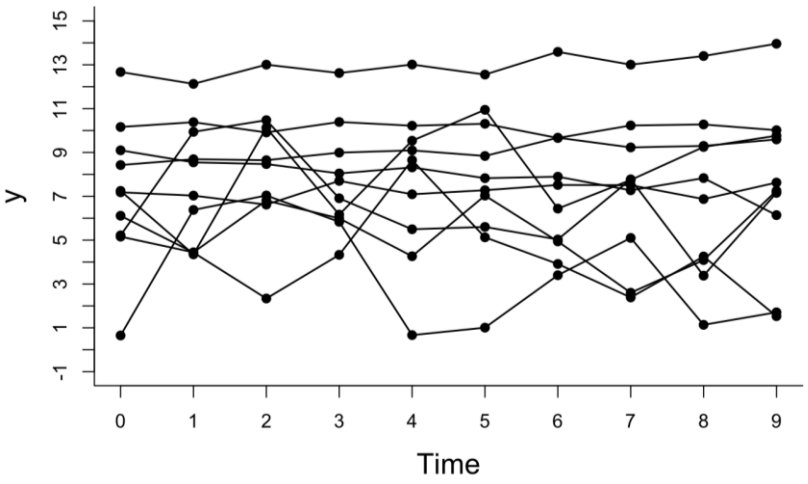
To visualize the concept of random variability, Figure 1-1 illustrates the hypothetical scenarios of randomly varying intraindividual variability and intragroup variability. As shown in Figure 1-1a, some individuals are more stable over repeated measurements while some others are more erratic from measurement to measurement. Similarly, Figure 1-1b shows that some clusters (e.g., cluster 3, 4) are more cohesive compared to the others (e.g., cluster 2, 11).

What is unobserved heterogeneity in random variability and why does it matter? In certain research scenarios, the population of interest is not homogeneous and does not follow one single distribution. It is not uncommon for the observed heterogeneity to be driven by unobserved sub-populations that are inherently distributional different. Often these subgroups are characterized by varying distributions in means, but this concept also naturally extends to situations where variability is randomly fluctuating. In such a case, the population may consist of several sub-populations that are different in the distribution of random variability (e.g., an unobserved subgroup of individuals may be highly unstable in their mental status compared to others; or an unobserved subgroup of organizations may demonstrate an exceptionally higher degree of cohesiveness relative to other organizations). Detecting such underlying heterogeneity in variability distribution holds clinical importance in various research areas. Intraindividual variability in cognitive functioning, for instance, is regarded as an important signal of potential future cognitive decline and dementia in older adults (Mumme et al., 2021). Therefore, there is

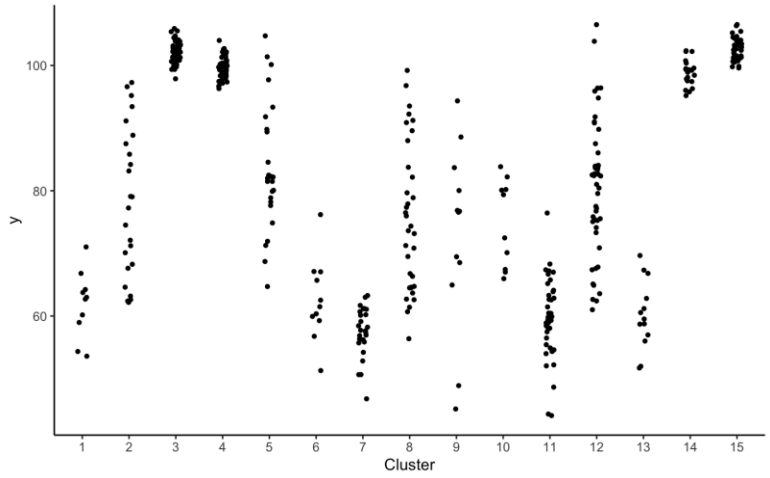
considerable clinical interest in early identification of individuals who may be more susceptible to age-related memory loss based on their distinctive intraindividual variability patterns.

Figure 1-1

Hypothetical Examples for Randomly Varying Intragroup Variability and Intraindividual Variability



(a)



(b)

Note. (a) Intragroup variability is randomly varying among different clusters; (b) Intraindividual variability is randomly varying across different individuals.

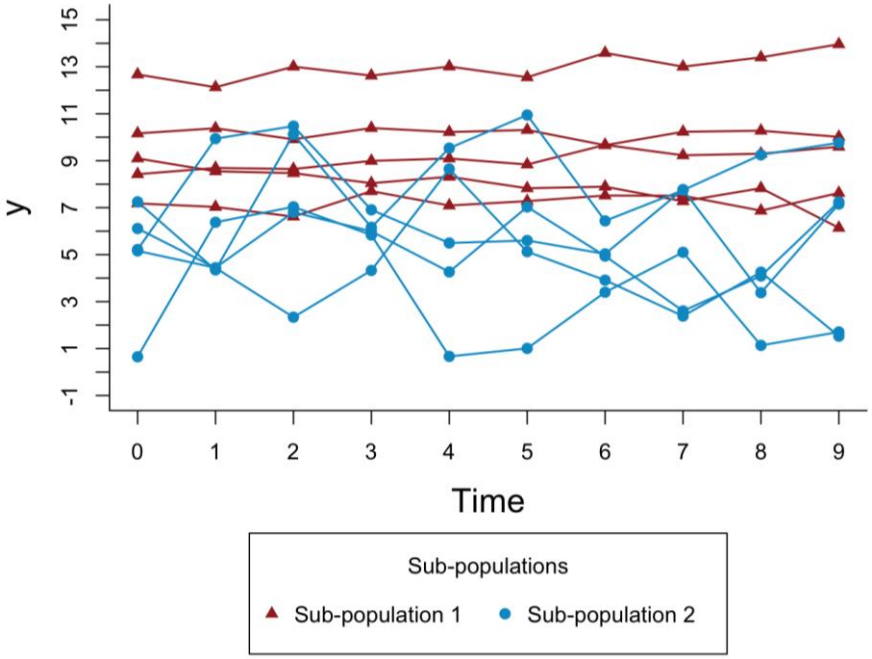
Additionally, regarding intragroup variability, organizations often grapple with the decision of fostering higher cohesion or encouraging greater diversity among their members. However, it is possible that there is no universal strategy that is uniformly applicable to all organizations—high cohesiveness may be advantageous for certain groups, while others may benefit from increased diversity. Detecting different types of groups based on such varying dynamics may thus be of great interest and importance for matters of policy considerations.

For the same hypothetical examples illustrated in Figure 1-1, Figure 1-2 shows the possible scenarios where the observed heterogeneity in variability is driven by the underlying sub-populations that differ in the patterns of random variability. In Figure 1-2a, there are two sub-populations—individuals belonging to latent class 1 are in general more consistent in the measured outcome *y*, while individuals belonging to latent class 2 have more erratic behavior patterns over time. Individuals in latent class 2 also demonstrate a greater level of diversity among themselves in terms of the level of intraindividual stability. In Figure 1-2b, there are three unobserved sub-populations—clusters belonging to latent class 3 are most homogeneous regarding the within-cluster distributions, while clusters belonging to latent class 1 show greater within-cluster variability. There is also a larger inter-cluster variation among the clusters in latent class 1 regarding the within-cluster cohesion.

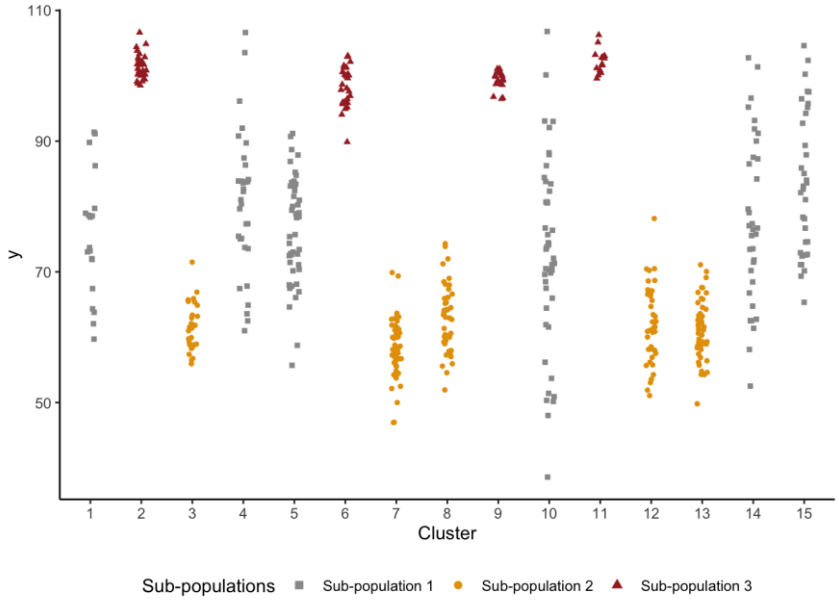
How should we study variability as the focal construct? For a very long time, applied researchers have been limited by available research methods that do not allow them to examine a wide range of theories about variability via empirical studies. Many of the familiar analytical models were developed to compare and model means, where variability is more of a nuisance parameter (e.g., t-test, ANOVA, regression). Although variability can be of important substantive interest, researchers often lack the proper analytical tools to model it as the focal

Figure 1-2

Hypothetical Examples Where There Are Unobserved Sub-populations That Differ in Randomly Varying Intragroup Variability and Intraindividual Variability



(a)



(b)

construct and make valid statistical inferences. Given the theoretical interests in variability, different methodological approaches have been proposed to study variability in the past decade, ranging from the more intuitive multi-stage approaches involving simple summary statistics (e.g., SD) to more sophisticated statistical models such as mixed effects location scale models (MELSM; Hedeker et al., 2008) and a multilevel structural equation modeling based approach (MSEM-RV; Feng & Hancock, 2022). These methodological advancements provide useful tools for researchers to model variability as a random coefficient that can be treated as the focal construct.

How should we detect unobserved heterogeneity in random variability? As variability modeling is a relative new topic in methodological literature, a huge challenge still remains when it comes to detecting the unobserved heterogeneity in random variability. Given that no existing research addresses the methodological considerations of uncovering the unobserved sub-populations that differ in intraindividual variability or intragroup variability, the current dissertation study aims to develop a model-based approach that can be employed to identify the underlying groups based on the distinct patterns of random variability.

1.2 Purpose of the Study

The primary purpose of the current study is to develop a latent variable based modeling framework to effectively detect unobserved sub-populations that vary in random intraindividual or intragroup variability. Importantly, the current study aims to address the research scenarios in which there are meaningful and theoretically justified unobserved subpopulations that drive the observed variation in within-level heteroscedasticity. Building upon the foundation established by previous work in MSEM-RV, the current study is conducted with the following specific aims:

1) Extend the previous work by introducing the theoretical framework of finite-mixture MSEM to model unobserved heterogeneity in random variability (MMSEM-RV) while taking measurement error into consideration.
2) Formulate and elucidate the specific estimation considerations for effectively implementing MMSEM-RV using a Bayesian approach, which includes the estimation procedure, software implementation, as well as results summary and interpretation.
3) Employ simulation studies to assess the performance of MMSEM-RV with Bayesian estimation across various conditions. More specifically, the simulation study aims to address the following research questions:
    a. Can MMSEM-RV be effectively estimated using a Bayesian approach via Gibbs sampling, yielding a converged solution?
    b. Can the model parameters of MMSEM-RV be accurately and efficiently estimated?
    c. How accurate is the latent class assignment when MMSEM-RV is used for classifying clusters, and how uncertain is the classification decision?
    d. To what extent do factors such as sample size, class separation, and measurement quality impact the estimation of MMSEM-RV?
    e. How does the proposed MMSEM-RV model compare to a naïve mean score based approach that ignore the measurement error?
4) Demonstrate the practical applicability of MMSEM-RV through a real-world data example.

1.3 Significance of the Study

The significance of this study is twofold, making substantial contributions to both applied research and methodological advancement for modeling variability as the key construct.

To begin with, the development of MMSEM-RV in this research holds relevance across a wide spectrum of scenarios where variability needs to be studied as the focal construct and when the need arises to identify sub-populations that are distributionally different with regard to variability. It thus has broad implications for applied research across many different fields. In clinical settings, for instance, it may be of substantial interest to identify different subgroups of individuals, especially those who are considered more vulnerable to a specific mental disorder. Notably, intraindividual inconsistency in cognitive functioning is considered an important signal of potential cognitive decline and future dementia among older adults (Mumme et al., 2021). To facilitate effective intervention for improved health outcomes, researchers and practitioners may wish to identify the individuals that are more susceptible to age-related memory loss based on the early patterns of intraindividual variability. Similarly, with organizational studies, it can be important to understand whether intragroup cohesion predicts group performance differently across different sub-populations. Detecting subgroups based on distinct intragroup dispersion patterns and their correlations with distal outcomes offers insights to refine strategies aimed at enhancing group productivity and performance.

On the other hand, aiming to address the broad research needs in detecting unobserved sub-populations based on patterns of *intraindividual variability* or *intragroup variability*, the proposed MMSEM-RV also contributes methodologically to the knowledge of advanced SEM, variability modeling, and mixture models. Although there exist several analytical approaches that can accommodate randomly varying variability, such as the MELSM (Hedeker & Mermelstein,

2007, 2012) and dynamic structural equation modeling (DSEM; Asparouhov et al., 2018; McNeish & Hamaker, 2020), as well as the more general MSEM-based random variability modeling framework (Feng & Hancock, 2020, 2022), no prior methodological endeavors have been dedicated to developing a finite-mixture model centered on variability as the central construct. The structure of such models as well as the appropriate estimation methods remain largely unexplored. Correspondingly, there are no known software packages with built-in functionality to estimate mixture random variability models. By formally defining the theoretical framework, articulating the estimation process of MMSEM-RV, evaluating the performance via simulation studies, as well as demonstrating the practical application using the real-world data example, this dissertation work offers an important methodological advancement in latent variable models of random variability.

1.4 Overview of the Chapters

The current chapter sets the stage by introducing the core concept of variability as a focal construct, as well as outlines the diverse research scenarios where population heterogeneity is driven by distinct sub-populations characterized by different patterns of random variability. It also establishes the groundwork for the subsequent chapters by presenting the rationale and objectives of the study. Chapter 2 (Literature Review) offers a thorough review of the existing methodological approaches for modeling variability and those for modeling the unobserved population heterogeneity in multilevel settings. The theoretical backgrounds for each approach are summarized and important limitations are discussed in this chapter. In Chapter 3 (Theoretical Framework), the conceptual framework of MMSEM-RV is introduced with great detail, along with an articulation of the model estimation procedures. Chapter 4 (Simulation Studies) presents

the design and implementation of the simulation studies. The simulation results are provided and summarized. Chapter 5 (Applied Illustrative Example) brings theory to practice by offering an applied illustrative example of MMSEM-RV using empirical data. Chapter 6 (Discussion) synthesizes the findings from theoretical development, simulation studies, and applied illustrative example. The chapter offers an in-depth discussion of the contribution of the study, the challenges remaining, the broader implications of MMSEM-RV framework, and the potential directions for future research.

Chapter 2. Literature Review

2.1 Modeling Random Variability

Over the past twenty years, methodological considerations for modeling variability have advanced substantially given the broad interest in variability across different fields. Some works are specifically motivated by the research need to investigate intraindividual variability in psychology studies, while some others address the variance heterogeneity more generally, in the context of educational settings, organizational research, ecological processes, or for the purpose of quality control. Although the various methods may originate from different perspectives, they share similar methodological grounds in many aspects. Based on their common theoretical foundations, the existing methods can be summarized into three categories, ranging from the more straightforward multi-step techniques that involve first estimating relevant summary statistics (e.g., standard deviation) as the input for further analysis, to more sophisticated statistical modeling frameworks that incorporate variability as a focal random coefficient, which include the multilevel modeling extensions (Blozis, 2022; Goldstein, 2011; Hedeker et al., 2008, 2012; Leckie et al., 2014; Lee & Nelder, 2006; Wang et al., 2012), and the more recently developed SEM based frameworks (Asparouhov et al., 2018; Feng & Hancock, 2020, 2022). This section reviews and summarizes the relevant findings from previous studies that focus on modeling variability.

2.1.1 Multi-step Approach

For study designs that follow one or more individuals over multiple repeated measurements, one popular approach that has been widely applied is to first compute a summary statistic across an individual's observed measures at different time points. This individual-level

summary statistic is designated to serve as the measure of intraindividual variability, which are then submitted to conventional statistical models for further analysis and inference. Typical examples of such summary statistics include intraindividual standard deviation (iSD), intraindividual mean square of successive differences (iMSSD), intraindividual residual standard deviation (iSDr), coefficient of variation (CV), and range (e.g., Adolph et al., 2015; Fagot et al., 2018; Hultsch & MacDonald, 2004; Ram et al., 2015; see Appendix A for more details). In addition to the basic summary statistics, other measures are also proposed to provide a more nuanced description of intraindividual variability. For example, Deboeck and colleagues (2009) discussed the use of derivative estimates to gauge intraindividual variability, which can subsequently be correlated with individual traits.

This multi-step approach is particularly useful for person-oriented analysis, which does not necessarily involve a population of interest. However, despite its easy implementation, such an approach has its own limitations. Importantly, the summary statistic conflates meaningful intraindividual fluctuations with measurement error, which is not properly accommodated in subsequent analyses. Further, the inherent uncertainties in estimating the statistics from sample data are also ignored when using the summary statistics as though they were true variabilities, which can lead to increased false positive rates in follow-up statistical tests (Dzubur et al., 2020).

2.1.2 Mixed Effects Modeling Approaches

When the research design involves repeated measures within individuals (or multiple units nested within groups), and that the observed individuals (or groups) are a random sample from a population of key inferential interest, researchers often utilize multilevel modeling (MLM; also known as *hierarchical linear modeling* or *mixed effects models*) to model such nested data structures. MLM partitions the variance into components at different levels (e.g.,

within-person vs. between-person, within-group vs. between-group) and thus inferences can be made at each level accordingly (see, e.g., Hox et al., 2017; Raudenbush & Bryk, 2002; Snijders & Bosker, 2011). A basic two-level MLM usually takes the form:

$$Y_{ij} = \mathbf{X}'_{ij}\boldsymbol{\beta} + \mathbf{Z}'_{ij}\mathbf{u}_j + e_{ij} \ , \tag{2.1}$$

where Y_{ij} is the observed outcome variable for level-1 unit i nested within level-2 cluster j; $\mathbf{X}_{ij}$ is the design matrix for fixed effects $\boldsymbol{\beta}$ and $\mathbf{Z}_{ij}$ is the design matrix[1] for random effects $\mathbf{u}_j$; e_{ij} is the level-1 residual, which is usually assumed to have a constant variance across all the level-2 units in conventional MLM: $e_{ij} \sim \text{N}(0, \sigma^2)$. In other words, it assumes all the individuals have the same level of stability, or all the groups have the same level of cohesiveness, after conditioning on the relevant covariates.

In order to accommodate variance heterogeneity, several extensions of conventional MLM have been proposed. Some models extend the conventional MLM by incorporating a structured model for the level-1 variance as a function of predictors. For example, in their textbooks, Goldstein (2011, p. 75 and p. 203), Raudenbush and Bryk (2002, p. 131), and Snijders and Bosker (2012, pp. 119-127) discussed possible specifications of the additional model for level-1 variance, which can be a linear, quadratic, or log linear function of the relevant predictors. Nelder and Lee (1991) proposed an extension of MLM called *generalized linear models* that incorporate a structured model for the level-1 dispersion parameter, which is applicable to a wide range of distributions (also see Lee & Nelder, 2001). These models aim to relax the strict assumption of constant variance and account for variance heterogeneity, but it has

[1] For a random intercept model, the design matrix $\mathbf{Z}_{ij}$ only contains ones. For a random slope model, the design matrix contains ones in the first column and the values of other level-1 predictors (a subset of $\mathbf{X}_{ij}$) in subsequent columns that are allowed to have random slopes.

a major limitation in that the variance (or dispersion parameter) is only modeled as a deterministic function of the predictors—no random effects are allowed.

To further relax the homoscedasticity assumption and to model within-level variability as a quantity that can randomly vary across level-2 units, further extensions of the MLM framework are proposed in the literature. Some of the earlier work treats the level-1 variance as a random variable by employing a two-stage approach: the level-1 residual variance estimates are first obtained via empirical Bayes, which are then regressed on a set of level-2 predictors as a single level analysis (Raudenbush & Bryk, 1987). Given the acknowledged drawbacks of the multi-step approaches, however, more effort was put into developing methodological approaches that can model the randomly varying mean and randomly varying dispersion simultaneously in one model. One possible solution is to model the heterogeneous variance as a random parameter that follows a certain distribution using a Bayesian approach. Kasim and Raudenbush (1998) proposed a method to model the heterogeneity of variance in two-level hierarchical models with a Gibbs sampler, where the within-level variance is specified to follow an inverse χ^2 distribution (and as a result varying randomly across the level-2 units).

On the other hand, other work along these lines has tried to build a structured component into the modeling of randomly varying within-level variance. Representative examples include the following: 1) the double hierarchical generalized linear models (DHGLM), where random effects can be specified in both the models for mean and dispersion parameter separately (Lee et al., 2017, pp. 137-140; Lee & Nelder, 2006); 2) the dynamic multilevel first-order autoregressive model for intensive longitudinal data (ILD), where random effects are allowed in the models for the autoregressive parameter (*inertia*) and the within-level variability of residuals (*innovation*), both of which indicate intraindividual stability (Jongerling et al., 2015; Wang et al., 2012); and

3) the mixed effects location scale model (MELSM) that was originally developed for analysis of ecological momentary assessment (EMA) data, which allows random effects in modeling both the means (*location*) and within-subject variability (*scale*) (Hedeker et al., 2008; Hedeker & Mermelstein, 2007, 2012; Leckie et al., 2014).

In a nutshell, these independently developed methodological approaches all extend the conventional MLM framework by allowing for the modeling of the within-level variability as a function of covariates and random effects. The DHGLM models the dispersion parameter φ for the residuals e_{ij} as (Lee et al., 2017, pp. 137-140; Lee & Nelder, 2006):

$$g(\varphi) = \mathbf{X}'\boldsymbol{\beta} + \mathbf{Z}'\mathbf{v} \tag{2.2}$$

where $g(\cdot)$ is a link function and $\mathbf{v}$ denotes the random effects in dispersion. For linear mixed effects models, the residuals are usually assumed to follow a normal distribution and as such the dispersion parameter φ is the residual variance σ_j^2. On the other hand, the dynamic multilevel first-order autoregressive model specifies a series of random effects models for the autoregressive parameter ϕ and residual variance σ_j^2 as (Jongerling et al., 2015; Wang et al., 2012):

$$\begin{aligned} \log(\sigma_j^2) &= \beta_0 z_j + \zeta_{0,j} \\ \phi_{j,1(K)} &= \beta_1 z_j + \zeta_{1,j} \\ &\cdots \\ \phi_{j,K(K)} &= \beta_K z_j + \zeta_{K,j} \end{aligned} , \tag{2.3}$$

where z_j is a selected predictor for person j and $\phi_{j,k(K)}$ ($k = 1, 2, \ldots, K$) is the kth order autoregressive parameter for person j that quantifies the intraindividual variability due to temporal dependency. The error terms $\zeta_{k,j}$ ($k = 0, 1, \ldots, K$) denote the random effects for each corresponding parameter.

Similarly, drawing upon the earlier research work on modeling variance heterogeneity in single-level regression (Aitkin, 1987) and the modeling of random location and scale effects (Cleveland et al., 2002), MELSM incorporates random effects for within-subject variance by specifying the log-linear model (Hedeker et al., 2008; Hedeker & Mermelstein, 2007, 2012):

$$\sigma^2_{e_{ij}} = \exp(\mathbf{W}'_{ij}\boldsymbol{\tau} + u_{1j}) \ , \tag{2.4}$$

where $\mathbf{W}_{ij}$ denotes the covariates that predict the within-cluster variability, and u_{1j} is the random effect of the within-level variance after controlling for the covariates. Importantly, within this modeling framework, u_{1j} is allowed to covary with other level-2 random effects $\mathbf{u}_j$ for the means. For example, it is possible to assess whether there is any linear relation between individuals' long-run average (or starting level) of the observed outcome and intraindividual fluctuation over repeated measures. Researchers can further allow the level-2 covariance components differ across level-2 units, either by directly modeling the covariance with a separate function (Leckie et al., 2014) or by indirectly allowing the corresponding standard deviations to vary (e.g., Rast & Ferrer, 2018; Williams et al., 2021). Most commonly, the LSM is used for ILD, such as ecological momentary assessment (EMA) data (e.g., Hedeker et al., 2008), although it has also been shown to be applicable for heterogeneous level-1 variance in a more generalized two-level scenario (e.g., Leckie et al., 2014).

The aforementioned MLM extensions offer useful analytical tools when the outcome variable of interest is directly observable and can be accurately measured without measurement error (or only with reasonably minimal negligible measurement error, such as heart rate). They can be applied to relax the often unrealistic assumption about constant variance, account for heterogeneous within-level variability, and investigate what factors can explain the heterogeneity in variability.

The MLM-based approaches are not without limitations, however, particularly in the social sciences and educational studies, where researchers are often interested in latent constructs that are not directly observable but can only be indicated by multiple error-laden observed variables. For instance, psychologists may wish to study the true underlying depression, but they can only collect observable symptom data through indirect measures such as surveys of internal feelings, observations of external behaviors, and/or relevant biomarkers and brain images. Similarly, education researchers interested in students' true math ability rely on the observed responses to a set of test items to make inference. To put it simply, observed variables rarely perfect measures of a latent construct and often contain some measurement error. Treating an observed variable as if it perfectly reflects the construct of interest can yield biased results, reduce statistical power, and lead to misleading conclusions (Cole & Preacher, 2014). Unfortunately, MLM and its extensions typically limit analyses to observed variables or a linear composite thereof (e.g., an arithmetic mean or sum score). As straightforward as it may seem, the sum score in fact relies on a highly constrained measurement model, which is not warranted without theoretical support and empirical validation (D. McNeish & Wolf, 2020). All the aforementioned models, however, simply ignore the measurement error altogether. Only recently, Blozis (2022) extended the MELSM by incorporating a common factor model to account for the measurement error in the outcome with longitudinal repeated measures.

In addition to the issue of measurement error, the various MLM extensions for heterogenous variability is also limited in other aspects. The existing literature has focused primarily on a single observed outcome, with only very few extensions to multivariate case (e.g., Kapur et al., 2015, specifically for intensive longitudinal data). Further, when both the mean and variability are being modeled simultaneously, this modeling framework only allows for the

modeling of within-level variability as an outcome. In some research scenarios, however, it may be important to investigate whether variability predicts other distal outcomes. For example, a gerontologist can be interested in whether the consistency in seniors' short-term memory predicts their degree of self-sufficiency with regard to life-care skills. To address these type of questions, a multi-stage approach is proposed for MELSM: the random effects are first imputed based on the model parameter estimates obtained at the first stage, which are then carried over to the next-stage analyses as predictors (Dzubur et al., 2020). It still, however, remains challenging to directly model variability as a random variable while serving as an outcome, predictor, or both simultaneously in one single model.

Additionally, the MLM-based approaches may not be able to provide a satisfactory solution in more complex research scenarios that are not typically considered under this framework. One of the potential scenarios involves comparison between observed groupings of the level-2 units (e.g., female vs. male, young adults vs. senior adults, public schools vs. private schools) not only in terms of the average means and within-level variability, but also in terms of the covariance structure among their corresponding random effects. Following the MLM convention, if the covariance components are each modeled as a separate function of the grouping covariates, it can easily result in nonpositive definiteness. To address this issue, a separation strategy has been proposed in the literature that decomposes the covariance matrix into standard deviations and correlations that can be modeled separately (Barnard et al., 2000). MELSM adopts this approach in covariance modeling, by incorporating covariates in the log-linear model for standard deviations (e.g., Rast & Ferrer, 2018; Williams et al., 2021). Although this approach ensures the positive definiteness of the covariance matrix, it assumes a constant correlation matrix, which may be another unrealistic assumption in multi-group analysis.

2.1.3 Multilevel Structural Equation Modeling Approaches

SEM is a multivariate analytical framework that has wide applications across various fields, including the social, behavioral, and education sciences. It is highly versatile and thus can be applied in many research scenarios. Importantly, it can model hypothesized structural and non-causal links among not just observed variables, but also latent constructs that are free of measurement error by incorporating measurement models as part of the structural equations.

Although historically SEM and MLM were developed as two distinct modeling frameworks, SEM and MLM actually share a lot of common ground (Curran, 2003). In MLM, unobserved latent variables are introduced as the random effects to account for the dependence among the observations nested within same higher-level unit. In SEM, unobserved latent variables are introduced as the common factors to account for the interdependence among observed indicators measuring the same constructs (Rabe-Hesketh et al., 2004). Given the analytical similarity between SEM and MLM, research efforts have been made to integrate MLM and SEM into a more general methodological framework that can accommodate population heterogeneity, clustering data structure, measurement error, and multivariate outcomes simultaneously (Goldstein & McDonald, 1988; Heck & Reid, 2023; McDonald & Goldstein, 1989; Muthén, 1989, 1994; Rabe-Hesketh et al., 2004). Multilevel SEM (MSEM) is thus developed as a synthesis of MLM and SEM, with performance comparable to traditional MLM but with the many versatility advantages of SEM (Bauer, 2003; Curran, 2003). Within MSEM, random effects can be conveniently modeled as latent factors, and as such they can be embedded within a more general latent variable modeling framework, where a wide variety of structural and nonstructural relations can be possibly examined. Further, MSEM does not assume perfect measures. It allows for measurement models at different levels, for both the outcomes and predictors to account for measurement error.

Following the within-between framework notations (e.g., Asparouhov & Muthen, 2007; Heck & Reid, 2023; Rabe-Hesketh et al., 2012), the general form of two-level MSEM can be summarized in the following equations:

$$Y_{pij} = Y_{pj}^{(\mathrm{B})*} + Y_{pij}^{(\mathrm{W})*} \tag{2.5}$$

$$\mathbf{Y}_{ij}^{(\mathrm{W})*} = \boldsymbol{\Lambda}^{(\mathrm{W})}\boldsymbol{\eta}_{ij}^{(\mathrm{W})} + \boldsymbol{\varepsilon}_{ij}^{(\mathrm{W})} \tag{2.6}$$

$$\boldsymbol{\eta}_{ij}^{(\mathrm{W})} = \mathbf{B}^{(\mathrm{W})}\boldsymbol{\eta}^{(\mathrm{W})}{}_{ij} + \boldsymbol{\Gamma}^{(\mathrm{W})}\mathbf{X}_{ij}^{(\mathrm{W})} + \xi_{ij}^{(\mathrm{W})} \tag{2.7}$$

$$\mathbf{Y}_{j}^{(\mathrm{B})*} = \boldsymbol{\nu}^{(\mathrm{B})} + \boldsymbol{\Lambda}^{(\mathrm{B})}\boldsymbol{\eta}_{j}^{(\mathrm{B})} + \boldsymbol{\varepsilon}_{j}^{(\mathrm{B})} \tag{2.8}$$

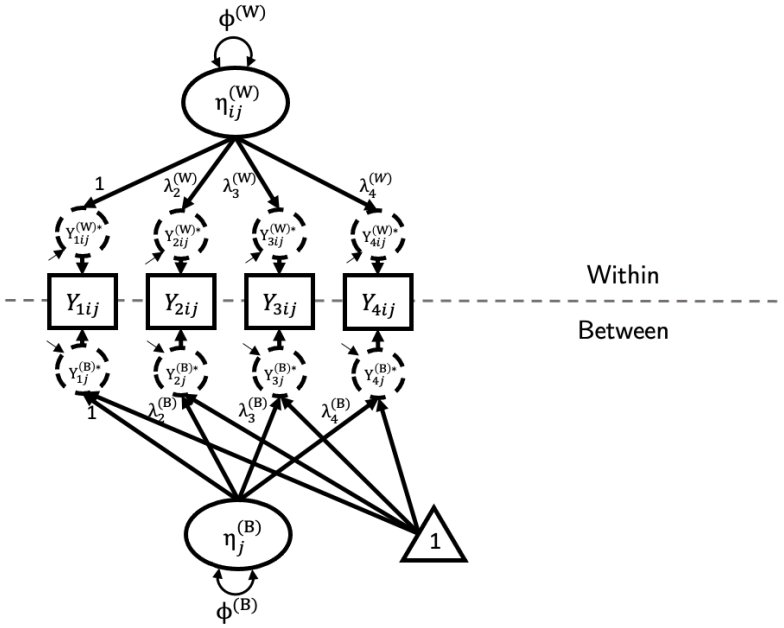
$$\boldsymbol{\eta}_{j}^{(\mathrm{B})} = \boldsymbol{\alpha}^{(\mathrm{B})} + \mathbf{B}^{(\mathrm{B})}\boldsymbol{\eta}_{j}^{(\mathrm{B})} + \boldsymbol{\Gamma}^{(\mathrm{B})}\mathbf{X}_{j}^{(\mathrm{B})} + \xi_{j}^{(\mathrm{B})} \ , \tag{2.9}$$

where Y_{pij} is the p^{th} observed indicator for lower-level unit i belonging to higher-level unit j. The observed variable is modeled as the sum of latent variables at the between-cluster level ($Y_{pj}^{(\mathrm{B})*}$, i.e., random intercept) and the within-cluster level ($Y_{pij}^{(\mathrm{W})*}$, i.e., deviation from cluster intercept). The measurement model at the within level is expressed in Eq. (2.6), where $\boldsymbol{\eta}_{pj}^{(\mathrm{W})}$ is a vector containing the latent constructs defined at the within level (e.g., individual's depressive mood assessed daily, or each team member's unobserved attitudes toward hustle culture). The structural model at the within level is given by Eq. (2.7), where $\mathbf{X}_{ij}^{(\mathrm{W})}$ denotes the observed covariates at the within level, and $\boldsymbol{\xi}_{ij}^{(\mathrm{W})}$ denotes the within-level random residuals for the latent constructs. Similarly, the measurement model and the structural model at the between level are given in Eqs. (2.8)-(2.9), respectively, where $\boldsymbol{\nu}^{(\mathrm{B})}$ is a vector of grand means and $\boldsymbol{\eta}_{j}^{(\mathrm{B})}$ is a vector containing the latent constructs defined at the between level for the higher-level unit j (e.g., an individual's long-term depressiveness, organizational level endorsement of hustle culture). In the MSEM literature, the residual vectors $\boldsymbol{\varepsilon}_{ij}^{(\mathrm{W})}$, $\boldsymbol{\xi}_{ij}^{(\mathrm{W})}$, $\boldsymbol{\varepsilon}_{j}^{(\mathrm{B})}$, and $\boldsymbol{\xi}_{j}^{(\mathrm{B})}$ are assumed to contain

independent normally distributed random variables with means of zero and the corresponding variance-covariance matrices $\boldsymbol{\Theta}^{(W)}$, $\boldsymbol{\Psi}^{(W)}$, $\boldsymbol{\Theta}^{(B)}$, and $\boldsymbol{\Psi}^{(B)}$ (Asparouhov & Muthén, 2007; Heck & Reid, 2023; Rabe-Hesketh et al., 2012). As suggested by the subscripts, the within-level variance-covariance $\boldsymbol{\Psi}^{(W)}$ are conventionally assumed to be invariant across clusters. A path diagram for a two-level SEM with one common latent factor at both levels is provided in Figure 2-1.

Figure 2-1

A Path Diagram for a Two-level MSEM Model



Note. For each within-level unit, four observed measures ($Y_1 - Y_4$) are collected as indicators of the latent construct.

Recent advancements in SEM, however, have seen the development of different approaches that relax the restricted assumption of constant within-level variances. For example, as part of a multilevel confirmatory factor analysis, Stapleton and colleagues (2016) discussed the use of a *phantom variable*—an unmeasured variable following a standard normal distribution

with no measured indicators—for this purpose. The phantom variable is introduced as a predictor of the within-level construct with a path that varies at the between level, thereby permitting the *scale* of within-cluster variability to be modeled as a random coefficient. This strategy can be applied to study both intraindividual variability and intragroup variability, although caution needs to be exercised in interpreting the scale parameter without further positivity constraints.

Another relevant recent development is dynamic structural equation modeling (DSEM), designed to model the dynamics of stable process longitudinal data. DSEM integrates the methodological techniques of MLM, SEM, time-varying effects modeling (TVEM), and time series analysis, where the current outcomes are modeled to be dependent on past outcomes (Asparouhov et al., 2018). DSEM is specifically developed for modeling longitudinal data that are measured for a large number of times within a short time period (e.g., EMA data and daily diary data). Importantly, by incorporating the log-transformed within-level residual variance as a random parameter, DSEM allows for individually-varying and time-varying within-level variability. Although DSEM can well accommodate intraindividual variability as a latent random variable within the SEM framework, the current theoretical developments have focused more on the structural model of modeling the variability as an outcome, and less as a predictor. The current software implementation of DSEM is available in M*plus* (Muthén & Muthén, 1998-2023). It is currently limited to two-level models using Bayesian estimation and its performance relies on large sample size (Schultzberg & Muthén, 2018).

Besides DSEM, some other lines of research in recent years also have emerged to take advantage of the structured latent variable modeling techniques when modeling heterogeneous intraindividual variability. They combine the techniques of MELSM and SEM, or more specifically, confirmatory factor analysis (CFA), to better estimate intraindividual variability.

Blozis (2022), for example, proposed a latent variable MELSM, where a common factor model is incorporated into MELSM to model the measurement structure of the longitudinal single-dimensional outcome; the common factor, which is free of measurement error, rather than the error-laden observed variable, is then modeled to follow the MELSM specification. With this approach, it is the latent factor's residual variance that is modeled as a function of covariates with the corresponding random effects. Nestler (2020, 2021), on the other hand, proposed a confirmatory factor level and variability model (CFLVM) for multivariate outcomes. In CFLVM, the intraindividual residual variances across multiple outcomes are assumed to load on common latent variability factor(s), thereby taking into account the reliability of the variability measure for each observed variable. The latent variability factor(s) can then be examined to make inferences in terms of the mean and variance-covariance parameters. Although both approaches focus specifically on intraindividual variability with ILD, Blozis' and Nestler's work exemplifies the flexibility of SEM framework, which has a great potential to offer in conjunction with the MLM-based approach.

At the same time, the MSEM based framework for modeling random variability (MSEM-RV) proposed by Feng and Hancock (2020, 2022) offers a general approach to modeling within-level variability. In MSEM-RV, the variability (both intraindividual and intragroup variability) is modeled as the focal latent construct, which has the flexibility to be an outcome, predictor, or both that is embedded within a broader covariance structure. This framework is different from the mixed effects modeling based approaches (e.g., MELSM, DHGLM) as it takes into account the measurement error of the observed indicators and it explicitly models the random within-level variability as a latent variable that goes beyond merely being the outcome variable—it can also be the predictor of other distal outcomes. MSEM-RV is also different from the other

previous or contemporary SEM-based approaches (e.g., DSEM, CFLVM) as it has a broader scope than merely intraindividual variability with ILD. Although MSEM-RV may appear to be similar to CFLVM, they reflect fundamentally different theoretical beliefs about the data generating process and imply distinct covariance structures. MSEM-RV is essentially built upon MSEM, where the within-level latent constructs and between-level latent constructs are substantively meaningful and of theoretical interest (see Stapleton et al., 2016). The random variability is then introduced to describe the dispersion of this theoretically meaningful within-level latent construct that is free of measurement error. In contrast, CFLVM does not involve any meaningful latent constructs at the within-level. Additionally, they also have different implications for the covariance structure among observed variables. With CFLVM, the within-level residuals of the observed variables are independent across different measures, after the random latent variability factor(s) are taken into account. The structural models of MSEM-RV, on the other hand, still allow the within-level residuals across different observed measures to covary after conditioning on the between-level random effects, which can be further explained by theoretically meaningful within-level latent construct(s).

The MSEM-RV framework has shown to be useful in a variety of scenarios. It can be applied in situations when multiple observed measures are collected at each measurement occasion to indicate common latent constructs (e.g., daily positive affect) or when multiple observed measures are administered to each sampled group member to measure a meaningful latent construct at the individual level (e.g., personal perceived support). Notably, to account for the heterogeneous within-level variability between observed sub-populations, MSEM-RV provides a viable solution to not only comparing the average variability, but also the covariance structure involving variability. It can be easily implemented using the multigroup analysis

paradigm of SEM, where the non-positive definiteness issues are much better circumvented. Given these important considerations, the current dissertation project chooses to further extend MSEM-RV to investigate unobserved sub-populations with regard to variability.

The basic form of a two-level MSEM-RV is briefly presented here (Figure 2-2), with more details available in Feng and Hancock (2020, 2022). The following model can be applied to study a continuous outcome with repeated measures nested within individuals, or multiple observations nested within a group. For simplicity all the models are presented for a two-level scenario, and it is assumed that the outcome measures are observed for individuals nested under groups to avoid unnecessary verbosity. But the same model can also be applied to repeated measures just as well.

The MSEM-RV has the multilevel measurement structure similar to what is specified in MSEM. In a case where there exists a unidimensional latent construct at both levels, the structural model for the p^{th} observed indicator for individual i in group j can be expressed as:

$$Y_{pij} = \gamma_{p00} + \lambda_p^{(B)}\xi_j^{(B)} + \varepsilon_{pj}^{(B)} + \lambda_p^{(W)}\xi_{ij}^{(W)} + \varepsilon_{pij}^{(W)}\ , \tag{2.10}$$

where γ_{p00} is the grand mean, $\xi_j^{(B)}$ denotes the between-level latent construct, and $\xi_{ij}^{(W)}$ denotes the within-level latent construct. Most commonly, the factor loadings $\lambda_p^{(B)}$ and $\lambda_p^{(W)}$ should be constrained to be equal (i.e., *cross-level measurement invariance*) for meaningful interpretation of the between-level construct (for more details refer to Jak, 2014, 2019; Jak & Jorgensen, 2017; Kim et al., 2018; Muthén, 1989; Stapleton et al., 2016). Alternatively, it is also possible to have a saturated structure at the between level if the between-level latent construct is not of inferential interest or not interpretable (see Stapleton et al., 2016). Eq. (2.10) can be equivalently expressed in the compact matrix form:

$$\mathbf{Y} = \boldsymbol{\Gamma} + \boldsymbol{\Lambda}^{(B)}\boldsymbol{\xi}^{(B)} + \boldsymbol{\varepsilon}^{(B)} + \boldsymbol{\Lambda}^{(W)}\boldsymbol{\xi}^{(W)} + \boldsymbol{\varepsilon}^{(W)}. \tag{2.11}$$

The within-level residuals ($\boldsymbol{\varepsilon}^{(W)}$) and between-level residuals ($\boldsymbol{\varepsilon}^{(B)}$) for the observed indicators are assumed to be independently and normally distributed: $\boldsymbol{\varepsilon}^{(W)} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Theta}^{(W)})$, $\boldsymbol{\varepsilon}^{(B)} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Theta}^{(B)})$. The within-level latent construct $\xi_{ij}^{(W)}$ is assumed to be normally distributed given variance σ_j^2, $\xi_{ij}^{(W)} \mid \sigma_j^2 \sim N(0, \sigma_j^2)$. The variance of the within-level latent construct for cluster *j* is denoted as σ_j^2, indicating it is cluster-specific. One common parameterization is to model the natural logarithm (ln) of σ_j^2 as a random latent variable at between-level[2]. The between-level structural equations can thus be written as

$$\begin{bmatrix} \xi_j^{(B)} \\ \ln(\sigma_j^2) \end{bmatrix} = \begin{bmatrix} 0 \\ \gamma_{10} \end{bmatrix} + \begin{bmatrix} \nu_{0j} \\ \nu_{1j} \end{bmatrix},$$

or in following equivalent matrix form[3]:

$$\boldsymbol{\xi}_j^{*(B)} = \boldsymbol{\gamma}^{(B)} + \mathbf{v}_j^{(B)}, \tag{2.12}$$

where $\boldsymbol{\xi}_j^{*(B)}$ is a vector of between-level latent coefficients containing both the between-level latent construct and the log of within-level variance; $\boldsymbol{\gamma}^{(B)}$ is the vector of fixed effects, and $\mathbf{v}_j^{(B)}$ contains the between level random effects. The level-2 random effects $\mathbf{v}_j^{(B)}$ are assumed to be multivariate normal:

$$\begin{aligned} \mathbf{v}_j^{(B)} &\sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Phi}_v) \\ \boldsymbol{\Phi}_v &= \begin{bmatrix} \phi_0 & \\ \phi_{10} & \phi_1 \end{bmatrix} \end{aligned} \tag{2.13}$$

[2] From the modeling perspective, the *random variability* is not considered as entirely stochastically random. With MMSEM-RV, it can be modeled as systematically varying while involving stochastic random errors.

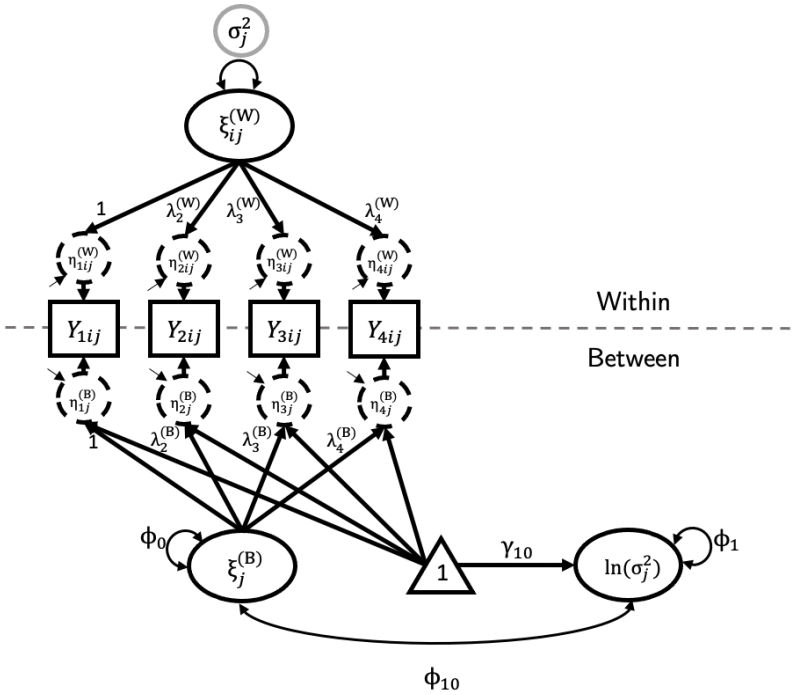
[3] Although not explicitly elaborated here for the sake of simplicity, the between-level structural model can be expanded to include structural relations among the level-2 random effects.

where ϕ_1 represents the between-level variance of the log residual variance, and ϕ_{10} denotes the covariance between the random intercepts and random log variances. Therefore, the model implies that the level-2 latent variables (random effects) follow a multivariate normal distribution, given the mean vector $\boldsymbol{\gamma}^{(B)}$ and covariance matrix $\boldsymbol{\Phi}_v$:

$$\xi_j^{*(B)} \mid \boldsymbol{\gamma}^{(B)}, \Phi_v \sim N\left(\boldsymbol{\gamma}^{(B)}, \Phi_v\right) . \tag{2.14}$$

Figure 2-2

MSEM-RV with One Within-Level Outcome Variable and One Between-Level Latent Outcome Variable



Note. For each within-level unit, four observed measures ($Y_1 - Y_4$) are collected as indicators of the latent construct. $\xi^{(W)}$ = within-level latent factor; $\xi^{(B)}$ = between-level latent factor.

2.2 Model-based Approaches for Unobserved Population Heterogeneity

2.2.1 Finite Mixture Models

Various methods have been proposed in the literature to address unobserved population heterogeneity. The goal, in brief, is to identify smaller and more homogeneous subgroups of the heterogeneous population. It often also involves classifying the individual observations into one of the sub-populations (i.e., clustering). Besides the numerous nonparametric algorithms for clustering, one popular model-based approach is finite mixture modeling (FMM). Within the FMM framework, it is assumed that the population consists of several sub-populations (latent classes). In each latent class, the data are modeled to follow a class-specific distribution. The probability density function (or probability mass function in discrete cases) of an observation from a finite mixture distribution has the general form of

$$f(\mathbf{y};\Theta) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{y};\vartheta_k), \tag{2.15}$$

where K represents the total number of latent classes and π_k denotes the mixing proportion for the k^{th} latent class, which is the probability that an observation belongs to the k^{th} sub-population. By definition, $0 \le \pi_k \le 1 \ \forall\ k = 1, 2, \ldots, K$ and $\sum_{k=1}^{K} \pi_k = 1$. In each latent class, the data follow a density distribution denoted by $f_k(\mathbf{y}; \vartheta_k)$, with class-specific parameters ϑ_k. The functional form of $f_k(\cdot)$ is usually known and prespecified. The unknown model parameters $\boldsymbol{\Theta}$ include the mixing proportions $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_K)'$ and class-specific distribution parameters $\boldsymbol{\vartheta} = (\vartheta'_1, \vartheta'_2, ..., \vartheta'_K)'$. Although it relies on proper parametric assumptions, one advantage of FMM compared to other nonparametric options of clustering is that it can model the latent class membership as a function of relevant covariates, where the predictive relations can be formally tested for statistical

inference. The latent class can also be modeled and tested as the predictor for distal outcomes, either within one single analytical model or via multi-step approaches (Nylund-Gibson et al., 2019).

2.2.2 Multilevel Mixture Models

To accommodate multilevel data structure in finite mixture modeling, multilevel mixture models have been proposed and discussed in the methodological literature (Asparouhov & Muthén, 2008; Muthén & Asparouhov, 2009; Vermunt, 2008, 2011; Vermunt & Magidson, 2005). Multilevel mixture models are also often referred to as *multilevel regression mixtures* because it has a close structural resemblance to the multilevel regression models, where the focal observed outcome is specified as a function of exogenous predictors. On the other hand, similar as FMM, multilevel mixture models assume the population is made of smaller homogeneous sub-populations by modeling it with latent categorical variables. Given the multilevel data structure, however, the clustering can occur at different levels. Therefore, in multilevel mixture models, the latent classes can be specified at within level only (L1MIX), between level only (L2MIX), or both within level and between level simultaneously (L1L2MIX; see more details in Asparouhov & Muthén, 2008; Sterba & Rights, 2022). It is also possible to include only categorical latent classes but no normally distributed random effects (e.g., Vermunt, 2011). The interpretation and model implication thereby differ accordingly under different parameterization specifications. The multilevel mixture models with latent random effects and latent classes specified at between level (L2MIX) are briefly summarized here as it is most relevant for the purpose of the current study. The L2MIX model for an observed outcome y is specified as:

$$
\begin{aligned}
y_{ij}\Big|_{C_j=k} &= \mathbf{X}_{ij}^{(\mathrm{W})}{}'\boldsymbol{\beta}_k^{(\mathrm{W})} + \mathbf{X}_j^{(\mathrm{B})}{}'\boldsymbol{\beta}_k^{(\mathrm{B})} + \mathbf{Z}_{ij}{}'\mathbf{u}_j + \varepsilon_{ij} \\
\varepsilon_{ij} &\sim \mathrm{N}\,(0, \sigma_k^2) \\
\mathbf{u}_j &\sim \mathrm{N}\,(0, \mathbf{T}_k) \qquad\qquad , \qquad\qquad (2.16) \\
P\left(C_j = k\right) = \pi_{jk} &= \frac{\exp(\alpha_k + \mathbf{X}_j{}'\boldsymbol{\Gamma}_k)}{\sum_k \exp(\alpha_k + \mathbf{X}_j{}'\boldsymbol{\Gamma}_k)}
\end{aligned}
$$

where *i* indexes the within-level units (e.g., individual students within schools or daily measures of a person) and *j* indexes the higher-level units. The between-level latent class membership for cluster *j* is denoted by C_j, which can take on values of $k = 1, 2, \ldots, K$. As in FMM, the data are allowed to follow a class-specific distribution. Therefore, the outcome y_{ij} is modeled conditional on the class membership C_j, usually specified as a function of fixed effects and random effects. The vectors $\mathbf{X}_{ij}^{(\mathrm{W})}$ and $\mathbf{X}_j^{(\mathrm{B})}$ contain the within-level and between-level covariates (including a constant 1 for fixed intercepts), respectively. The vector $\mathbf{Z}_{ij}$ contains the subset within-level predictors that have random slopes (in addition to a constant 1 for random intercepts). The fixed effect coefficients $\boldsymbol{\beta}_k^{(\mathrm{W})}$ for within-level predictors and fixed effect coefficients $\boldsymbol{\beta}_k^{(\mathrm{B})}$ for between-level predictors are allowed to vary across latent classes, as suggested by the subscript *k*. The within-level residual ε_{ij} follows a normal distribution with variance σ_k^2, which varies across latent classes. The random effects $\mathbf{u}_j$ follow a multivariate normal distribution with covariance matrix $\mathbf{T}_k$, which is also allowed to vary across the *K* latent classes. Finally, the probability for cluster *j* belonging to latent class *k*, π_{jk}, is usually modeled using a multinomial logistic function, where $\mathbf{X}_j$ is between-level covariates that can predict the latent class membership.

The L2MIX multilevel mixture models partially address the unobserved population heterogeneity in within-level variability, by allowing the within-level residual variance to take on different values in clusters that belong to different latent classes. The limitation of this approach,

however, is that all clusters are constrained to have constant within-level variability given the latent class membership. It also inherits other limitations of mixed effects models, an important one being that measurement error is not taken into consideration in the modeling process.

2.2.3 Finite Mixture Multilevel Structural Equation Modeling

Given the limitations of the multilevel regression mixtures on one hand, and the capability of SEM framework to flexibly model error-free latent constructs, it is a natural direction for researchers to consider combing them into a more powerful analytical tool that enjoys the advantages of both approaches. Mixture multilevel structural equation modeling was thus proposed, which incorporates measurement model, multilevel modeling, and mixture modeling into a unified framework (see more details in Asparouhov & Muthén, 2008; Kim et al., 2016). Similar to multilevel mixture models, mixture MSEM allows for the specification of latent classes at different levels. The choice of parameterization again needs to be justified by the corresponding theoretical beliefs. Assuming the theory suggests the presence of latent classes at the between level (e.g., heterogeneity in student performance is due to the different teaching styles adopted by teachers), a mixture MSEM with between-level latent classes is thus an appropriate model to consider. Methodologically speaking, the mixture MSEM with between-level latent classes provides greater flexibility in the between-level structural model—all model parameters, including the measurement model at both levels and the residual variance-covariance matrix, can be modeled as class-specific (Asparouhov & Muthén, 2008).

In this section, a basic mixture MSEM with between-level latent classes (i.e., *between-class multilevel factor mixture model*; Kim et al., 2016) is presented and briefly reviewed as it has direct implications for the proposed modeling approach in the following chapter. Assuming a

multilevel data structure with continuous observed indicators used to measure continuous latent constructs, the model is specified as follows:

$$
\begin{aligned}
\mathbf{Y}_{ij}\Big|_{C_j=k} &= \boldsymbol{\nu}_k + \boldsymbol{\Lambda}_k^{(\mathrm{B})}\boldsymbol{\eta}_{jk}^{(\mathrm{B})} + \boldsymbol{\Lambda}_k^{(\mathrm{W})}\boldsymbol{\eta}_{ijk}^{(\mathrm{W})} + \boldsymbol{\varepsilon}_{jk}^{(\mathrm{B})} + \boldsymbol{\varepsilon}_{ijk}^{(\mathrm{W})} \\
\boldsymbol{\varepsilon}_{jk}^{(\mathrm{B})} &\sim \mathrm{N}\left(0, \Theta_k^{(\mathrm{B})}\right) \\
\boldsymbol{\varepsilon}_{ijk}^{(\mathrm{W})} &\sim \mathrm{N}\left(0, \Theta_k^{(\mathrm{W})}\right) \\
\boldsymbol{\eta}_{jk}^{(\mathrm{B})} &\sim \mathrm{N}\left(\kappa_k^{(\mathrm{B})}, \Phi_k^{(\mathrm{B})}\right) \\
\boldsymbol{\eta}_{ijk}^{(\mathrm{W})} &\sim \mathrm{N}\left(\kappa_k^{(\mathrm{W})}, \Phi_k^{(\mathrm{W})}\right) \\
P(C_j = k) = \pi_{jk} &= \frac{\exp(\alpha_k + \mathbf{X}_j'\boldsymbol{\Gamma}_k)}{\sum_k \exp(\alpha_k + \mathbf{X}_j'\boldsymbol{\Gamma}_k)}
\end{aligned}
\quad , \qquad (2.17)
$$

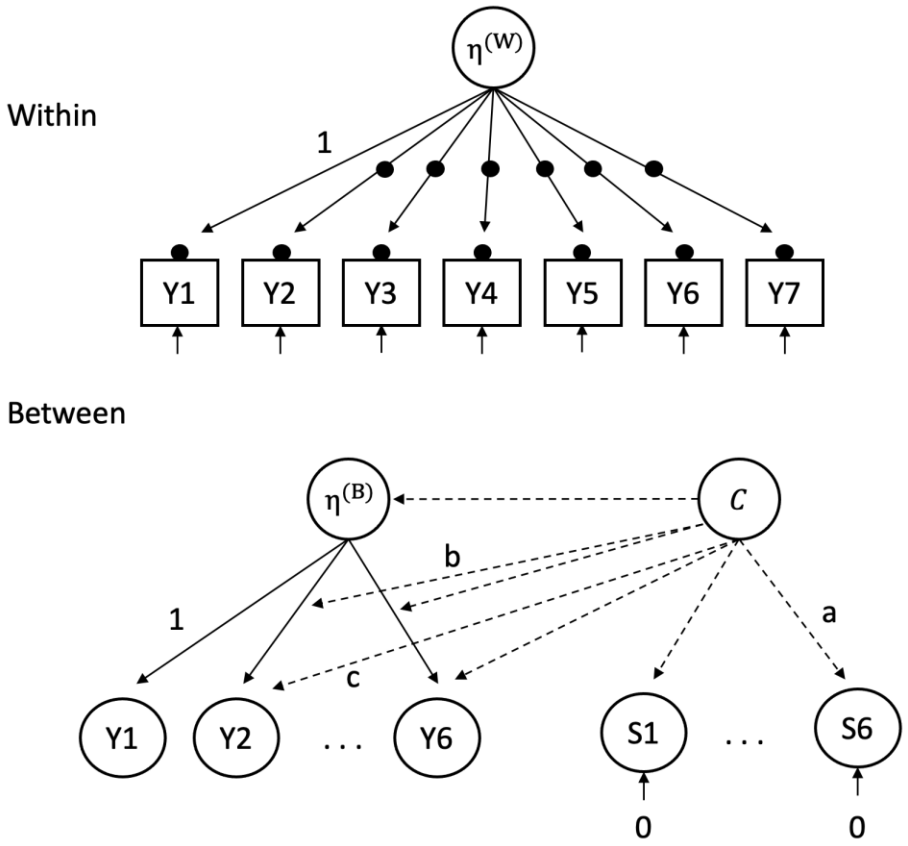
where $\mathbf{Y}_{ij}$ is a vector containing the observed measures for lower-level unit *i* belonging to the higher-level unit *j*. The between-level factor loadings $\boldsymbol{\Lambda}_k^{(\mathrm{B})}$ and within-level factor loadings $\boldsymbol{\Lambda}_k^{(\mathrm{W})}$ are class specific, allowing for unique measurement models in each latent class. The residuals at the within-level ($\boldsymbol{\varepsilon}_{ijk}^{(\mathrm{W})}$) and between-level ($\boldsymbol{\varepsilon}_{jk}^{(\mathrm{B})}$) are both assumed to be normally distributed with class-specific covariance. The latent constructs at the within-level ($\boldsymbol{\eta}_{ijk}^{(\mathrm{W})}$) and between-level ($\boldsymbol{\eta}_{jk}^{(\mathrm{B})}$) are also assumed to be normally distributed with class-specific parameters. The residuals are independent of the latent continuous variables on both levels, and the residuals at different levels are independent of each other: $\eta_k \perp \varepsilon_k$, $\boldsymbol{\varepsilon}^{(\mathrm{B})} \perp \boldsymbol{\varepsilon}^{(\mathrm{W})}$. Finally, the between-level latent class membership is modeled as a function of the cluster-level covariates with a multinomial logistic model.

Built upon MSEM, mixture MSEM explicitly incorporates measurement models in both levels without assuming perfect observed measures. Similar to multilevel mixture models, mixture MSEM also partially addresses the heterogeneity in within-level variability by allowing the variance of within-level latent constructs ($\Phi_k^{(\mathrm{W})}$) to vary among latent classes. But again, the

within-level variance can only take on discrete values and it is assumed to be constant within each latent class.

Figure 2-3

Path Diagram for Between-Class Multilevel Factor Mixture Model



Note. The solid dots indicate random coefficients at the between level (e.g., random intercepts and within-level factor loadings). The dashed arrows with letters a, b, and c are used to denote class-specific within-level factor loadings, between-factor loadings, and between-level intercept fixed effects, respectively. S2–S6 = random within-level factor loadings; $\eta^{(W)}$ = within-level continuous common latent factor; $\eta^{(B)}$ = between-level continuous common latent factor; *C* = between-level latent classes. Adapted from Kim et al. (2016).

A path diagram for a two-level factor mixture model with between-level latent classes is presented in Figure 2-3 (Kim et al., 2016). As shown in the figure, in the mixture models literature, latent class membership is commonly visualized as a latent (categorical) variable

represented by a circle. Directed paths are drawn from the latent categorical variable to each of the parameters that are allowed to vary across the latent classes. Despite the popularity of this convention, the diagram can get increasingly confusing very quickly as the model grows in complexity. Therefore, in the current study, the multiple group analysis paradigm is instead employed to visualize the mixture multilevel structural eqaution models with random variability.

Chapter 3. Theoretical Framework

3.1 Finite Mixture MSEM for Random Variability

This study aims to propose and investigate a mixture MSEM framework for random variability (MMSEM-RV). It integrates mixture MSEM and MSEM-RV to offer greater flexibility in modeling random variability. The proposed modeling approach allows for both continuous random effects and latent categorical variables in explaining the heterogeneity in within-level variability. It has a broad appeal as it can be applied to study both variability over repeated measures and dispersion within groups. Given the limitations of existing analytical tools, MMSEM-RV is an important addition to the methodological literature on modeling variability.

Let vector $\mathbf{Y}_{ij}$ denote the observed outcome measures for lower-level unit i in higher-level cluster j. The between-level latent class membership for cluster j is denoted by C_j. With MMSEM-RV, the conditional multilevel measurement model for $\mathbf{Y}_{ij}$ given the latent class C_j ($C_j = 1, 2, \ldots, k$) is the following[4]:

$$\mathbf{Y}_{ij}\Big|_{C_j=k} = \boldsymbol{\alpha}_k + \boldsymbol{\Lambda}_k^{(\mathrm{B})}\boldsymbol{\xi}_{jk}^{(\mathrm{B})} + \boldsymbol{\varepsilon}_{jk}^{(\mathrm{B})} + \boldsymbol{\Lambda}_k^{(\mathrm{W})}\boldsymbol{\xi}_{ijk}^{(\mathrm{W})} + \boldsymbol{\varepsilon}_{ijk}^{(\mathrm{W})}, \tag{3.1}$$

where $\boldsymbol{\alpha}_k$ is a vector containing the grand intercepts for the observed measures over all the clusters within latent class k. The within-level latent constructs and factor loadings are denoted by $\boldsymbol{\xi}_{ijk}^{(\mathrm{W})}$ and $\boldsymbol{\Lambda}_k^{(\mathrm{W})}$, respectively. The between-level latent constructs and factor loadings are denoted by $\boldsymbol{\xi}_{jk}^{(\mathrm{B})}$ and $\boldsymbol{\Lambda}_k^{(\mathrm{B})}$, respectively, assuming there are meaningful between-level constructs of theoretical interest. Otherwise, a saturated covariance structure can be specified at the between-

[4] Eq. (3.1) can be interpreted as the data generating model in a given latent class k. However, the latent class membership C_j for each cluster is unobserved and thus will be estimated with uncertainty.

level. Importantly, to interpret $\boldsymbol{\xi}_{jk}^{(\mathrm{B})}$ as the configural construct at the aggregated level, $\boldsymbol{\Lambda}_{k}^{(\mathrm{W})}$ and $\boldsymbol{\Lambda}_{k}^{(\mathrm{B})}$ should be constrained equal (Jak, 2019; Jak & Jorgensen, 2017; Stapleton et al., 2016). The within-level and between-level residuals are each denoted by $\boldsymbol{\varepsilon}_{ijk}^{(\mathrm{W})}$ and $\boldsymbol{\varepsilon}_{jk}^{(\mathrm{B})}$, respectively, which are assumed to be multivariate normally distributed while being independent of each other and independent of the latent constructs at both levels:

$$\begin{aligned}
&\boldsymbol{\varepsilon}_{jk}^{(\mathrm{B})} \sim \mathrm{N}\left(0, \boldsymbol{\Theta}_{k}^{(\mathrm{B})}\right) \\
&\boldsymbol{\varepsilon}_{ijk}^{(\mathrm{W})} \sim \mathrm{N}\left(0, \boldsymbol{\Theta}_{k}^{(\mathrm{W})}\right) \\
&\boldsymbol{\xi}_{k}^{(\mathrm{B})} \perp \boldsymbol{\varepsilon}_{k}^{(\mathrm{B})} \\
&\boldsymbol{\xi}_{jk}^{(\mathrm{W})} \perp \boldsymbol{\varepsilon}_{jk}^{(\mathrm{W})} \\
&\boldsymbol{\varepsilon}_{k}^{(\mathrm{B})} \perp \boldsymbol{\varepsilon}_{jk}^{(\mathrm{W})}
\end{aligned} \quad . \tag{3.2}$$

The within-level residual variance-covariance is denoted by $\boldsymbol{\Theta}_{k}^{(\mathrm{W})}$ and between-level residual variance-covariance is denoted by $\boldsymbol{\Theta}_{k}^{(\mathrm{B})}$, both allowed to vary across the latent classes. Depending on the measurement structure, there can be multiple latent constructs at the within level and thus $\boldsymbol{\xi}_{ijk}^{(\mathrm{W})}$ can be multidimensional. But for simplicity, for the current study let us assume $\boldsymbol{\xi}_{ijk}^{(\mathrm{W})}$ only contains one element $\xi_{ijk}^{(\mathrm{W})}$ that is of inferential interest. This within-level latent construct is assumed to be normally distributed with randomly varying variance σ_{jk}^{2}:

$$\xi_{ijk}^{(\mathrm{W})}\Big|_{\sigma_{jk}^{2}} \sim \mathrm{N}\left(0, \sigma_{jk}^{2}\right). \tag{3.3}$$

The structural model for between-level random coefficients $\boldsymbol{\xi}_{jk}^{*(\mathrm{B})} = (\xi_{jk}^{(\mathrm{B})}, \ln(\sigma_{jk}^{2}))'$ is:

$$\boldsymbol{\xi}_{jk}^{*(\mathrm{B})} = \boldsymbol{\gamma}_{k}^{(\mathrm{B})} + \boldsymbol{\Gamma}_{k}\mathbf{X}_{j} + \mathbf{v}_{jk}^{(\mathrm{B})}, \tag{3.4}$$

where $\boldsymbol{\gamma}_{k}^{(\mathrm{B})}$ is a vector containing the grand intercept for the random coefficients within latent class *k*. The random coefficients can be further modeled as a function of the between-level

covariates $\mathbf{X}_j$ with class-specific path coefficients $\mathbf{\Gamma}_k$. The random effects are denoted by $\mathbf{v}_{jk}^{(\mathrm{B})}$, which are assumed to be normally distributed with class-specific variance-covariance matrix:

$$\mathbf{v}_{jk}^{(\mathrm{B})} \sim \mathrm{N}\,(\mathbf{0}, \mathbf{\Phi}_{vk})\,. \tag{3.5}$$

Importantly, with MMSEM-RV, the within-level latent construct variability is modeled as a random coefficient that can vary across the clusters given the latent class. More specifically, within each latent class, the log-transformed within-level variance is modeled as a latent factor at the between-level. For a model with unidimensional between-level construct and unidimensional within-level construct, Eq. (3.4) is equivalent to

$$\begin{bmatrix} \xi_{jk}^{(\mathrm{B})} \\ \ln(\sigma_{jk}^2) \end{bmatrix} = \begin{bmatrix} 0 \\ \gamma_{10k} \end{bmatrix} + \begin{bmatrix} \gamma_{01k} & \cdots & \gamma_{0qk} \\ \gamma_{11k} & \cdots & \gamma_{1qk} \end{bmatrix} \begin{bmatrix} X_{1jk} \\ \vdots \\ X_{qjk} \end{bmatrix} + \begin{bmatrix} \nu_{0jk} \\ \nu_{1jk} \end{bmatrix}, \tag{3.6}$$

where the latent mean of $\xi_{jk}^{(\mathrm{B})}$ is constrained to zero for identification purpose. Eq. (3.5) can also be expressed as:

$$\begin{bmatrix} \nu_{0jk} \\ \nu_{1jk} \end{bmatrix} \sim \mathrm{N}\,(\mathbf{0}, \mathbf{\Phi}_{vk}) = \mathrm{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \phi_{0k} & \\ \phi_{10k} & \phi_{1k} \end{bmatrix} \right) \quad . \tag{3.7}$$

Finally, the latent class membership C_j can also be modeled as a function of the between-level covariates using the multinomial logistic link:

$$P(C_j = k) = \pi_{jk} = \frac{\exp(\tau_k + \mathbf{X}_j'\mathbf{B}_k)}{\sum_k \exp(\tau_k + \mathbf{X}_j'\mathbf{B}_k)}, \tag{3.8}$$

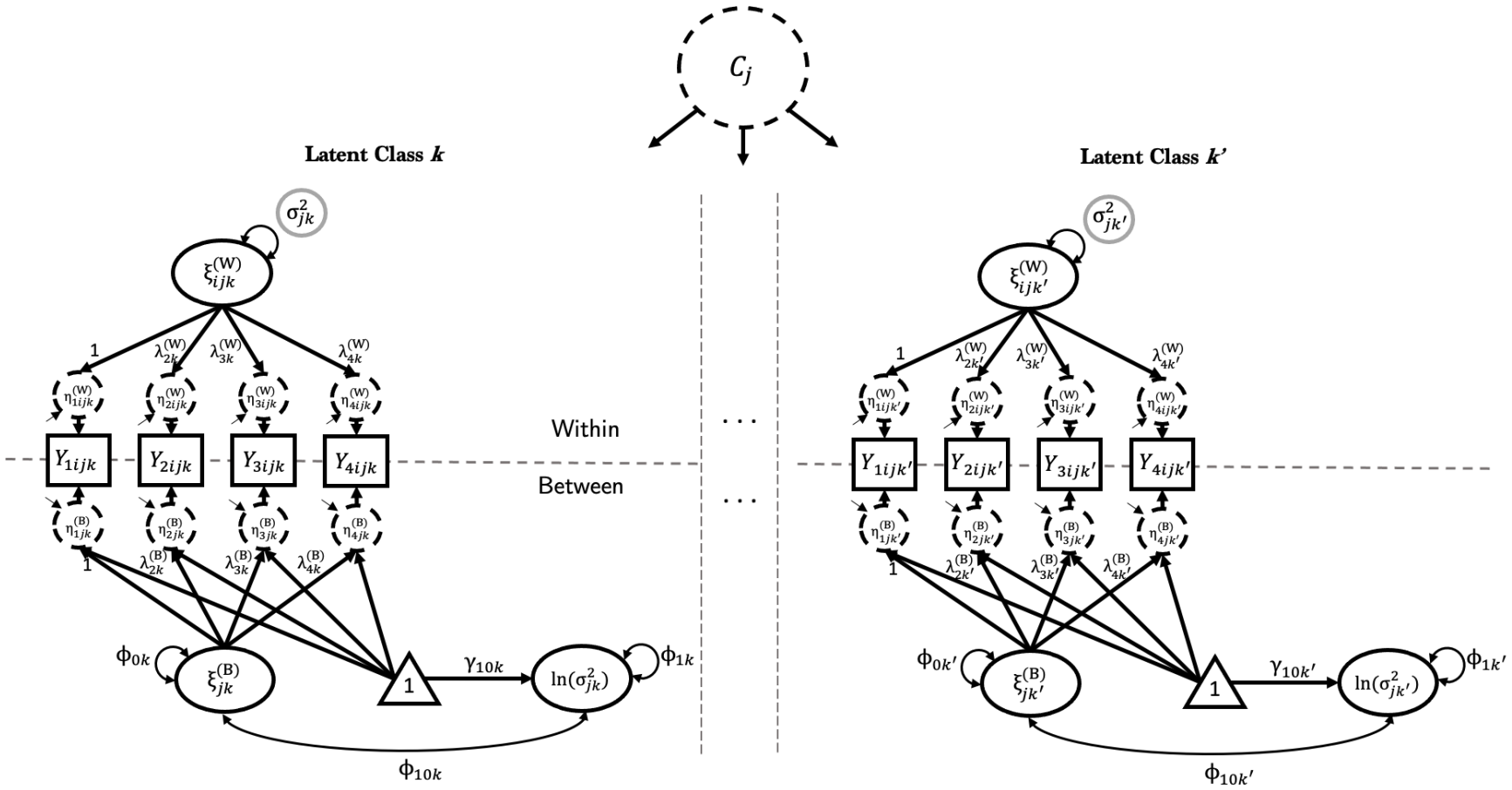
where τ_k denotes the class-specific intercept and $\mathbf{B}_k$ denotes the slopes. For model identification purposes, τ_1 and $\mathbf{B}_1$ are constrained to 0 with latent class 1 being treated as the reference class.

To illustrate, the path diagram for an unconditional MMSEM-RV with one between-level latent construct and one within-level latent construct is presented in Figure 3-1. The current

study will only focus on the unconditional model with a single latent construct at the within level. The extension to multidimensional latent constructs and conditional MMSEM-RV will be further considered in future research.

Figure 3-1

Path Diagram for MMSEM-RV with One Common Latent Factor at the Within-Level and the Corresponding Configural Construct at the Between-Level



Note. The latent class membership is denoted by $C_j = 1, 2, 3, \ldots, K$ and $k \neq k'$. To interpret $\xi^{(B)}_{jk}$ as the configural construct at the aggregated level, for each indicator the loadings $\lambda^{(W)}_{pk}$ and $\lambda^{(B)}_{pk}$ should be constrained equal.

3.2 Model Estimation

In practice, the population value of model parameters remains unknown and need to be estimated from sample data. In the literature, two estimation approaches are usually employed for models that allow for heterogeneous variability or unobserved population heterogeneity: maximum likelihood (ML) estimation and Bayesian estimation.

The univariate MELSM with one observed outcome variable has been traditionally estimated by marginal ML with Newton-Raphson algorithm using SAS `PROC NLMIXED` procedure (Hedeker et al., 2008) or the MIXREGLS program (Hedeker & Nordgren, 2013), where the random variability can only be modeled as an outcome. ML estimation is also used for the multivariate extension of MELSM with a latent variability factor, as detailed by Nestler (2020). A potential issue with the ML estimation approach for MELSM is that it is challenging to implement with small sample size (e.g., small number of clusters, few individuals, or measurement occasions), in which case the Bayesian framework offers a better solution as it does not rely on large sample theories. Therefore, in practice, the Bayesian approach via Markov Chain Monte Carlo (MCMC) simulation is commonly employed for estimating univariate (Rast et al., 2012) as well as multivariate longitudinal MELSM (Kapur et al., 2015) in JAGS/WinBUGS. Recent developments in MELSM for univariate observed outcomes further propose a two-stage approach that combines ML and Bayesian estimation. It utilizes the empirical Bayes methods to obtain multiple draws of the random coefficients based on the stage-1 ML parameter estimates, which can in turn be used in stage-2 statistical analyses as predictors in the MixWILD program (Dzubur et al., 2020).

On the other hand, in the MSEM literature, although ML and Bayesian estimation have both been employed to estimate MSEM models (Asparouhov & Muthén, 2016), Bayesian

estimation is a more common choice as it is easier and more straightforward to implement given the complexity of such models. As MSEM models usually do not have a closed-form likelihood function, the ML estimation needs to approximate the likelihood function by numerically integrating over all the random effects and latent variables, which can be quite computationally demanding as the number of unobserved latent factors increases. Particularly, when measurement models are present, Bayesian approach serves as a popular alternative. Bayesian approach is also shown to be able to overcome the convergence problem with ML and improve the accuracy of parameter estimates for MSEM (Depaoli & Clifton, 2015). The latest version of the popular SEM software M*plus* (version 8.9; Muthén & Muthén, 1998-2023) offers built-in DSEM/MSEM modules that can model random residual variance on its log-transformed scale via Bayesian estimation, which greatly facilitates the implementation of random variability models in practice. Although ML is traditionally used to estimate MELSM, fitting the MELSM as a MSEM or DSEM via Bayesian estimation has been shown to be straightforward and computationally efficient (e.g., McNeish, 2021; Nestler, 2021).

With regard to finite mixture models and mixture multilevel models, conventionally the ML estimation via Expectation-Maximization (EM) algorithm is most commonly used in practice (e.g., Asparouhov & Muthén, 2008). Its dominance is facilitated by the availability of various software tools, such as M*plus*, that provide built-in ML estimation for mixture models. However, in recent years the Bayesian approach has also been discussed as a compelling alternative, which not only has greater stability in estimates but also offers the advantage of accommodating an unspecified number of classes (e.g. Bainter, 2015).

Given the anticipated increasing model complexity as well as the corresponding computational burden for latent variable models in the context of modeling random variability,

Bayesian estimation was proposed for MSEM-RV and was shown to be effective across various modeling scenarios (Feng & Hancock, 2022). Therefore, the current proposed study also chooses to employ and further assess the performance of Bayesian estimation for the mixture extension of MSEM-RV.

3.2.1 Bayesian Estimation

With Bayesian estimation, inference can be made about the model parameters using the posterior distribution given the observed data. As MMSEM-RV contains a collection of unobserved latent variables, we can instead work with the augmented posterior density given the observed data $\mathbf{Y}$:

$$P(\boldsymbol{\Theta} \mid \mathbf{Y}) \propto P(\mathbf{Y}, \boldsymbol{\Theta}), \tag{3.9}$$

where $\boldsymbol{\Theta}$ is a vector that contains the model parameters and the between-level latent random variables.

In the case of unconditional MMSEM-RV, given a known number of latent classes K, we have $\boldsymbol{\Theta} = (\boldsymbol{\theta}', \text{vec}(\boldsymbol{\xi}_1^{*(\text{B})})', \ldots, \text{vec}(\boldsymbol{\xi}_K^{*(\text{B})})', \boldsymbol{\varepsilon}_1^{(\text{B})'}, \ldots, \boldsymbol{\varepsilon}_K^{(\text{B})'})'$. Here the model parameters vector $\boldsymbol{\theta}$ contains the mixing proportions, class-specific factor loadings and intercepts, fixed effects, variance-covariance of the random effect, and the residual variance: $\boldsymbol{\theta} = (\boldsymbol{\pi}', \boldsymbol{\alpha}_1', \ldots, \boldsymbol{\alpha}_K', \boldsymbol{\Lambda}_1^{(\text{B})'}, \ldots, \boldsymbol{\Lambda}_K^{(\text{B})'}, \boldsymbol{\Lambda}_1^{(\text{W})'}, \ldots, \boldsymbol{\Lambda}_K^{(\text{W})'}, \boldsymbol{\gamma}_1^{(\text{B})'}, \ldots, \boldsymbol{\gamma}_K^{(\text{B})'}, \text{vech}(\boldsymbol{\Phi}_{\text{v}1})', \ldots, \text{vech}(\boldsymbol{\Phi}_{\text{v}K})', \text{vech}(\boldsymbol{\Theta}_1^{(\text{B})})', \ldots, \text{vech}(\boldsymbol{\Theta}_K^{(\text{B})})', \text{vech}(\boldsymbol{\Theta}_1^{(\text{W})})', \ldots, \text{vech}(\boldsymbol{\Theta}_K^{(\text{W})})')'$, where the mixing proportion vector $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_K)'$, with the constraint that $\Sigma\, \pi_K = 1$. The augmented posterior distribution can be obtained as Eq. (3.10) using Bayes rule:

$$P(\boldsymbol{\theta}, \boldsymbol{\xi}_1^{*(\text{B})}, \ldots, \boldsymbol{\xi}_K^{*(\text{B})}, \boldsymbol{\varepsilon}_1^{(\text{B})}, \ldots, \boldsymbol{\varepsilon}_K^{(\text{B})} \mid \mathbf{Y}) \propto P(\mathbf{Y}, \boldsymbol{\xi}_1^{*(\text{B})}, \ldots, \boldsymbol{\xi}_K^{*(\text{B})}, \boldsymbol{\varepsilon}_1^{(\text{B})}, \ldots, \boldsymbol{\varepsilon}_K^{(\text{B})} \mid \boldsymbol{\theta}) P(\boldsymbol{\theta}) \tag{3.10}$$

The first term on the right side of Eq. (3.10) is the complete data likelihood of the observed variables and between-level latent variables, which can be computed as follows assuming the observations are independent conditioning on cluster membership and the clusters are independent:

$$
\begin{aligned}
&P(\mathbf{Y}, \boldsymbol{\xi}_1^{*(\mathrm{B})}, \ldots, \boldsymbol{\xi}_K^{*(\mathrm{B})}, \boldsymbol{\varepsilon}_1^{(\mathrm{B})}, \ldots, \boldsymbol{\varepsilon}_K^{(\mathrm{B})} \mid \boldsymbol{\theta}) \\
&= \prod_{j=1}^{J} P(\mathbf{Y}_j, \boldsymbol{\xi}_1^{*(\mathrm{B})}, \ldots, \boldsymbol{\xi}_K^{*(\mathrm{B})}, \boldsymbol{\varepsilon}_1^{(\mathrm{B})}, \ldots, \boldsymbol{\varepsilon}_K^{(\mathrm{B})} \mid \boldsymbol{\theta}) \\
&= \prod_{j=1}^{J} \sum_{k=1}^{K} \pi_k \left\{ P(\mathbf{Y}_{jk}, \boldsymbol{\xi}_k^{*(\mathrm{B})}, \boldsymbol{\varepsilon}_k^{(\mathrm{B})} \left| \boldsymbol{\theta}_k^{*} \right.) \right\} \\
&= \prod_{j=1}^{J} \sum_{k=1}^{K} \pi_k \left\{ P(\mathbf{Y}_{jk} \left| \boldsymbol{\xi}_k^{*(\mathrm{B})}, \boldsymbol{\varepsilon}_k^{(\mathrm{B})}, \boldsymbol{\theta}_k^{*} \right.) P(\boldsymbol{\xi}_{jk}^{*(\mathrm{B})} \mid \boldsymbol{\theta}_k^{*}) P(\boldsymbol{\varepsilon}_{jk}^{(\mathrm{B})} \mid \boldsymbol{\theta}_k^{*}) \right\} \\
&= \prod_{j=1}^{J} \sum_{k=1}^{K} \pi_k \left\{ \prod_{i=1}^{n_j} \left[P(\mathbf{Y}_{ijk} \left| \boldsymbol{\xi}_{jk}^{*(\mathrm{B})}, \boldsymbol{\varepsilon}_{jk}^{(\mathrm{B})}, \boldsymbol{\theta}_k^{*} \right.) \right] P(\boldsymbol{\xi}_{jk}^{*(\mathrm{B})} \mid \boldsymbol{\theta}_k^{*}) P(\boldsymbol{\varepsilon}_{jk}^{(\mathrm{B})} \mid \boldsymbol{\theta}_k^{*}) \right\}
\end{aligned}
\quad , \tag{3.11}
$$

where $\boldsymbol{\theta}_k^{*}$ is a vector containing the class-specific model parameters: $\boldsymbol{\theta}_k^{*} = (\boldsymbol{\alpha}_k', \boldsymbol{\Lambda}_k^{(\mathrm{B})\prime}, \boldsymbol{\Lambda}_k^{(\mathrm{W})\prime}, \boldsymbol{\gamma}_k^{(\mathrm{B})\prime}$, vech($\boldsymbol{\Phi}_{\mathrm{v}k}$)′, vech($\boldsymbol{\Theta}_k^{(\mathrm{B})}$)′, vech($\boldsymbol{\Theta}_k^{(\mathrm{W})}$)′)′. Assuming multivariate normality and unidimensional within-level measurement model, the first term within the braces of Eq. (3.11) can be expanded as

$$
\begin{aligned}
&P(\mathbf{Y}_{ijk} \mid \boldsymbol{\xi}_{jk}^{*(\mathrm{B})}, \boldsymbol{\varepsilon}_{jk}^{(\mathrm{B})}, \boldsymbol{\theta}_k^{*}) \\
&= P(\mathbf{Y}_{ijk} \mid \boldsymbol{\xi}_{jk}^{*(\mathrm{B})}, \boldsymbol{\varepsilon}_{jk}^{(\mathrm{B})}, \boldsymbol{\alpha}_k, \boldsymbol{\Lambda}_k^{(\mathrm{B})}, \boldsymbol{\Lambda}_k^{(\mathrm{W})}, \boldsymbol{\gamma}_k^{(\mathrm{B})}, \boldsymbol{\Phi}_{\mathrm{v}k}, \boldsymbol{\Theta}_k^{(\mathrm{B})}, \boldsymbol{\Theta}_k^{(\mathrm{W})}) \\
&= \frac{1}{2\pi^{P/2} \mid \boldsymbol{\Lambda}_k^{(\mathrm{W})} \sigma_{jk}^2 \boldsymbol{\Lambda}_k^{(\mathrm{W})\prime} + \boldsymbol{\Theta}_k^{(\mathrm{W})} \mid^{1/2}} \\
&\exp\left\{ -\frac{1}{2} (\mathbf{Y}_{ijk} - (\boldsymbol{\alpha}_k + \boldsymbol{\Lambda}_k^{(\mathrm{B})} \boldsymbol{\xi}_{jk}^{(\mathrm{B})} + \boldsymbol{\varepsilon}_{jk}^{(\mathrm{B})}))' (\boldsymbol{\Lambda}_k^{(\mathrm{W})} \sigma_{jk}^2 \boldsymbol{\Lambda}_k^{(\mathrm{W})\prime} + \boldsymbol{\Theta}_k^{(\mathrm{W})})^{-1} (\mathbf{Y}_{ijk} - (\boldsymbol{\alpha}_k + \boldsymbol{\Lambda}_k^{(\mathrm{B})} \boldsymbol{\xi}_{jk}^{(\mathrm{B})} + \boldsymbol{\varepsilon}_{jk}^{(\mathrm{B})})) \right\}
\end{aligned}
\quad , \tag{3.12}
$$

with P denoting the number of observed indicators for the measurement model. The second and third terms within the braces of Eq. (3.11) constitute the conditional likelihood of between-level latent variables given model parameters. Based on Eqs. (3.4)-(3.7), the second term can be expressed as

$$P(\boldsymbol{\xi}_{jk}^{*(\mathrm{B})} \mid \boldsymbol{\theta}_k^*) = P(\boldsymbol{\xi}_{jk}^{*(\mathrm{B})} \mid \boldsymbol{\gamma}_k^{(\mathrm{B})}, \boldsymbol{\Phi}_{\nu k}) = \frac{1}{2\pi^{m/2} \mid \boldsymbol{\Phi}_{\nu k} \mid^{1/2}} \exp\left\{ -\frac{1}{2} (\boldsymbol{\xi}_{jk}^{*(\mathrm{B})} - \boldsymbol{\gamma}_k^{(\mathrm{B})})' \boldsymbol{\Phi}_{\nu k}^{-1} (\boldsymbol{\xi}_{jk}^{*(\mathrm{B})} - \boldsymbol{\gamma}_k^{(\mathrm{B})}) \right\} . \tag{3.13}$$

where m denotes the number of between-level random coefficients. In the case of the unconditional MMSEM-RV with unidimensional within-level measurement model, $m = 2$. Because the between-level residuals are assumed to follow a multivariate normal distribution, the third term within the braces of Eq. (3.11) can thus be expressed as

$$P(\boldsymbol{\varepsilon}_{jk}^{(\mathrm{B})} \mid \boldsymbol{\theta}_k^*) = P(\boldsymbol{\varepsilon}_{jk}^{(\mathrm{B})} \mid \boldsymbol{\Theta}_k^{(\mathrm{B})}) = \frac{1}{2\pi^{P/2} \mid \boldsymbol{\Theta}_k^{(\mathrm{B})} \mid^{1/2}} \exp\left\{ -\frac{1}{2} (\boldsymbol{\varepsilon}_{jk}^{(\mathrm{B})} - \mathbf{0})' (\boldsymbol{\Theta}_k^{(\mathrm{B})})^{-1} (\boldsymbol{\varepsilon}_{jk}^{(\mathrm{B})} - \mathbf{0}) \right\} . \tag{3.14}$$

The second term on the right side of Eq. (3.10) is the joint prior distribution of model parameters. The joint density of the model parameters can be further written in the following form assuming independence between the fixed effects and random effects:

$$\begin{aligned} & P(\boldsymbol{\theta}) \\ &= P(\boldsymbol{\pi}, \boldsymbol{\alpha}_1, \ \ldots, \boldsymbol{\alpha}_K, \boldsymbol{\Lambda}_1^{(\mathrm{B})}, \ldots, \boldsymbol{\Lambda}_K^{(\mathrm{B})}, \boldsymbol{\Lambda}_1^{(\mathrm{W})}, \ldots, \boldsymbol{\Lambda}_K^{(\mathrm{W})}, \boldsymbol{\gamma}_1^{(\mathrm{B})}, \ldots, \boldsymbol{\gamma}_K^{(\mathrm{B})}, \boldsymbol{\Phi}_{\nu 1}, \ldots, \boldsymbol{\Phi}_{\nu K}, \boldsymbol{\Theta}_1^{(\mathrm{B})}, \ldots, \boldsymbol{\Theta}_K^{(\mathrm{B})}, \boldsymbol{\Theta}_1^{(\mathrm{W})}, \ldots, \boldsymbol{\Theta}_K^{(\mathrm{W})}) \\ &= P(\boldsymbol{\pi}) P(\boldsymbol{\alpha}_1, \ \ldots, \boldsymbol{\alpha}_K) P(\boldsymbol{\Lambda}_1^{(\mathrm{B})}, \ldots, \boldsymbol{\Lambda}_K^{(\mathrm{B})}) P(\boldsymbol{\Lambda}_1^{(\mathrm{W})}, \ldots, \boldsymbol{\Lambda}_K^{(\mathrm{W})}) P(\boldsymbol{\gamma}_1^{(\mathrm{B})}, \ldots, \boldsymbol{\gamma}_K^{(\mathrm{B})}) P(\boldsymbol{\Phi}_{\nu 1}, \ldots, \boldsymbol{\Phi}_{\nu K}) \\ & P(\boldsymbol{\Theta}_1^{(\mathrm{B})}, \ldots, \boldsymbol{\Theta}_K^{(\mathrm{B})}) P(\boldsymbol{\Theta}_1^{(\mathrm{W})}, \ldots, \boldsymbol{\Theta}_K^{(\mathrm{W})}) \\ &= P(\boldsymbol{\pi}) \prod_k^K P_k(\boldsymbol{\alpha}_k) P_k(\boldsymbol{\Lambda}_k^{(\mathrm{B})}) P_k(\boldsymbol{\Lambda}_k^{(\mathrm{W})}) P_k(\boldsymbol{\gamma}_k^{(\mathrm{B})}) P_k(\boldsymbol{\Phi}_{\nu k}) P_k(\boldsymbol{\Theta}_k^{(\mathrm{B})}) P_k(\boldsymbol{\Theta}_k^{(\mathrm{W})}) \end{aligned} \tag{3.15}$$

Combining Eqs. (3.10)-(3.15), the augmented posterior distribution is the following:

$$
\begin{aligned}
&P(\boldsymbol{\theta}, \xi_1^{*(\mathrm{B})}, \ldots, \xi_K^{*(\mathrm{B})}, \boldsymbol{\varepsilon}_1^{(\mathrm{B})}, \ldots, \boldsymbol{\varepsilon}_K^{(\mathrm{B})} \mid \mathbf{Y}) \propto P(\mathbf{Y}, \xi_1^{*(\mathrm{B})}, \ldots, \xi_K^{*(\mathrm{B})}, \boldsymbol{\varepsilon}_1^{(\mathrm{B})}, \ldots, \boldsymbol{\varepsilon}_K^{(\mathrm{B})} \mid \boldsymbol{\theta}) P(\boldsymbol{\theta}) \\
&= \prod_{j=1}^{J} \sum_{k=1}^{K} \pi_k \left\{ \prod_{i=1}^{n_j} \left[P(\mathbf{Y}_{ijk} \left| \xi_{jk}^{*(\mathrm{B})}, \boldsymbol{\varepsilon}_{jk}^{(\mathrm{B})}, \boldsymbol{\theta}_k^{*} \right.) \right] P(\xi_{jk}^{*(\mathrm{B})} \mid \boldsymbol{\theta}_k^{*}) P(\boldsymbol{\varepsilon}_{jk}^{(\mathrm{B})} \mid \boldsymbol{\theta}_k^{*}) \right\} P(\boldsymbol{\theta}) \\
&= \prod_{j=1}^{J} \sum_{k=1}^{K} \pi_k \left\{ \begin{array}{l}
\left[\prod_{i=1}^{n_j} \frac{1}{2\pi^{P/2} \mid \boldsymbol{\Lambda}_k^{(\mathrm{W})} \sigma_{jk}^2 \boldsymbol{\Lambda}_k^{(\mathrm{W})\prime} + \boldsymbol{\Theta}_k^{(\mathrm{W})} \mid^{1/2}} \times \right. \\
\left. \exp\left\{ -\frac{1}{2} (\mathbf{Y}_{ijk} - (\boldsymbol{\alpha}_k + \boldsymbol{\Lambda}_k^{(\mathrm{B})} \xi_{jk}^{(\mathrm{B})} + \boldsymbol{\varepsilon}_{jk}^{(\mathrm{B})}))' (\boldsymbol{\Lambda}_k^{(\mathrm{W})} \sigma_{jk}^2 \boldsymbol{\Lambda}_k^{(\mathrm{W})\prime} + \boldsymbol{\Theta}_k^{(\mathrm{W})})^{-1} (\mathbf{Y}_{ijk} - (\boldsymbol{\alpha}_k + \boldsymbol{\Lambda}_k^{(\mathrm{B})} \xi_{jk}^{(\mathrm{B})} + \boldsymbol{\varepsilon}_{jk}^{(\mathrm{B})})) \right\} \right] \\
\times \frac{1}{2\pi^{m/2} \mid \Phi_{\nu k} \mid^{1/2}} \exp\left\{ -\frac{1}{2} (\xi_{jk}^{(\mathrm{B})} - \boldsymbol{\gamma}_k^{(\mathrm{B})})' \Phi_{\nu k}^{-1} (\xi_{jk}^{(\mathrm{B})} - \boldsymbol{\gamma}_k^{(\mathrm{B})}) \right\} \\
\times \frac{1}{2\pi^{P/2} \mid \boldsymbol{\Theta}_k^{(\mathrm{B})} \mid^{1/2}} \exp\left\{ -\frac{1}{2} (\boldsymbol{\varepsilon}_{jk}^{(\mathrm{B})} - \mathbf{0})' (\boldsymbol{\Theta}_k^{(\mathrm{B})})^{-1} (\boldsymbol{\varepsilon}_{jk}^{(\mathrm{B})} - \mathbf{0}) \right\}
\end{array} \right\} \\
&\quad \times P(\boldsymbol{\pi}) \prod_{k}^{K} P_k(\boldsymbol{\alpha}_k) P_k(\boldsymbol{\Lambda}_k^{(\mathrm{B})}) P_k(\boldsymbol{\Lambda}_k^{(\mathrm{W})}) P_k(\boldsymbol{\gamma}_k^{(\mathrm{B})}) P_k(\boldsymbol{\Phi}_{\nu k}) P_k(\boldsymbol{\Theta}_k^{(\mathrm{B})}) P_k(\boldsymbol{\Theta}_k^{(\mathrm{W})}) \quad .
\end{aligned}
$$

(3.16)

With the augmented posterior distribution, the parameter of focal interest can be estimated by examining its marginal posterior distributions via MCMC. For instance, the posterior mean can be used as parameter point estimate and a posterior credible interval can be obtained as the interval estimate. Bayesian estimation via MCMC can be implemented using Gibbs sampling in various software packages such as `rjags` (Plummer et al., 2022) or Hamiltonian Monte Carlo (HMC) sampling in `Stan` (Carpenter et al., 2017). For the illustrative examples of MSEM-RV, Gibbs sampling and HMC sampling yield similar results, while HMC sampling in `Stan` takes a substantially longer run-time compared to the Gibbs sampler in `JAGS` (Feng & Hancock, 2022). Therefore, the current proposed study will only use Gibbs sampler for model estimation.

3.2.2 Label Switching

Label switching is a common issue that often arises in the estimation of finite mixture models (Redner & Walker, 1984; Stephens, 2000). Because the latent classes are unobserved, the

label of the classes is therefore arbitrarily assigned. Given a fixed number of latent classes, the data likelihood is invariant across any permutations of the model parameters $\boldsymbol{\Theta}$ and the corresponding relabeling of the classes. Therefore, the latent class labels and class-specific parameters can be arbitrarily switched from replication to replication when the same model is fit to the data. For example, when two latent classes are enumerated, a latent class labeled as class 1 in one model fitting output can be labeled as class 2 in another replication, yielding the same likelihood.

The label switching problem becomes even more prominent with Bayesian estimation, as the class labels can permute from iteration to iteration; it is also likely to occur not only within the same MCMC chain, but also between the different chains (Cassiday et al., 2021; Papastamoulis, 2016; Rodríguez & Walker, 2014). If no prior information is available to better distinguish between the sub-populations, and the prior distributions are set the same for all permutations of $\boldsymbol{\Theta}$, the posterior distributions will be similarly symmetric, rendering the generated MCMC sample unidentifiable (Stephens, 2000). More specifically, the posterior marginal distributions will be highly similar—the posterior marginal classification probabilities for clustering any observation into each latent class will be $1/K$ and the posterior means for model parameters will be the same across all the classes, making them both misleading and useless for further interpretation (Rodríguez & Walker, 2014; Stephens, 2000). What further complicates the problem is that label switching is actually a desired condition for the MCMC algorithm to reach convergence, as the sampler should ideally explore and visit all the modes of the posterior distribution before convergence (Papastamoulis & Iliopoulos, 2010; Stephens, 2000).

Various methods are proposed to prevent the label switching problem (see Cassiday et al., 2021; Papastamoulis, 2016; Papastamoulis & Iliopoulos, 2010; Stephens, 2000). One of the most commonly-used approaches is to impose identifiability constraints on the parameter space. For example, if the ordering constraint is imposed on the intercepts of mixing proportions ($\tau_1 < \tau_2 < \tau_3 ... < \tau_K$) or on the fixed effects ($\gamma_{101} < \gamma_{102} < \gamma_{103} ... < \gamma_{10K}$) such that it can be satisfied by only one possible permutation of $\boldsymbol{\Theta}$, the label switching can be prevented throughout the iterations. This approach is straightforward to implement, although it requires a thorough prior knowledge of the sub-population distributions. Without a careful choice of the prior distributions or the identifiability constraints, the posterior distributions may still show multimodality (Stephens, 2000). The other commonly-implemented solution in practice is to limit the estimation to one Markov chain and set reasonable and well-separated initial values for the class-specific parameters. By doing so, one hopes that the chain stays within the neighborhood of one of the modes and the class labels will not switch from iteration to iteration within the chain. Some more advanced and principled solutions are also proposed to address the issue of label switching, including the relabeling algorithm (Stephens, 2000) and the Equivalence Classes Representatives algorithm (Papastamoulis & Iliopoulos, 2010), which involve an iterative process to identify an optimal permutation at each MCMC iteration. These approaches are particularly useful when the posterior distributions are genuinely multimodal. However, the major drawback of these iterative algorithms is that they are computationally demanding and require a great deal of memory storage.

Considering the model complexity, high dimensionality, and anticipated computational burden of estimating MMSEM-RV, this study chooses to employ the conventional approach that is computationally less demanding. This involves limiting the estimation process to a single

chain and applying identifiability constraints to prevent label switching. The application of more sophisticated and computationally intensive methods will be considered in future research.

3.2.3 Model Fit Assessment

In the current study, it is assumed that the MMSEM-RV model is correctly specified and thereby the model parameter estimates can be meaningfully interpreted upon convergence. But in practice model fit assessment is an important issue to consider. Particularly, given that MMSEM-RV consists of multiple components at multiple levels, there can be multiple potential sources of model misspecification. Model misspecification at any layer can adversely impact parameter estimation throughout the entire model. For example, given a misspecified measurement model at the within-level, the estimated fixed effects and random effects with regard to within-level variability can be misleading.

Recent developments in SEM offer useful discussions on global model fit assessment in multilevel settings for MSEM (Ryu, 2014; Ryu & West, 2009) as well as in single-level Bayesian SEM (Asparouhov & Muthén, 2021; Garnier-Villarreal & Jorgensen, 2020; Hoofs et al., 2018; Levy, 2011). In the case of MSEM-RV (Feng & Hancock, 2022) more specifically, a sequential approach is proposed such that the model (mis)fit is examined at different levels separately, which includes the measurement model at the within level, measurement model at the between level, and heterogeneous within-level variability structure. The first two sources of model misfit are assessed by comparing the saturated model to the corresponding reduced model with a restricted structure at either the within level or between level, via ML-based model fit indices (e.g., χ^2). The third piece of model misfit can be assessed by comparing the random variability model to the reduced model that constrained zero variation of the within-level variability, via an information criterion (e.g., BIC).

On the other hand, within the context of mixture SEM or mixture MSEM, the primary focus is on model comparison to determine the correct number of latent classes (Kim & Wang, 2017; Nylund-Gibson & Masyn, 2016; Tofighi & Enders, 2008), while nearly no study addresses the methodological considerations of global goodness of fit assessment or local fit assessment (with the only exception of Nagelkerke et al., 2016, where local fit indices were proposed for multilevel latent class models with categorical data). In summary, despite the collective efforts made to improve the evaluation of model fit in advanced latent variable modeling, the literature is still generally lacking for assessing model fit in mixture MSEM with Bayesian estimation. For the newly proposed MMSEM-RV more specifically, there is no existing guideline for model fit assessment that can be directly applied or adapted.

Although it is beyond the scope of the current study to develop a principled framework for model fit assessment in mixture MSEM or MMSEM-RV, it is an important issue that needs to be addressed in future research endeavors. Nonetheless, this study will demonstrate the practical application of information criterion-based model comparison for selecting the best fitting MMSEM-RV model via the real data illustrative example.

3.3 Measurement Invariance

In the current study, when multiple latent classes are enumerated, it is assumed that such unobserved population heterogeneity is driven only by the structural difference in variability or grand means, rather than measurement noninvariance—thereby setting a different scope from the multilevel factor mixture models (see Kim et al., 2016, 2017). Importantly, to meaningfully compare between latent classes in terms of the within-level variability, the following measurement invariance is assumed:

1) Configural invariance for within-level and between-level measurement models, that is, the measurement model at the within-level and between-level share the same factor structure. For a given unit *i* nested within cluster *j*, we have

$$\begin{matrix} \xi_{ijk}^{(\mathrm{B})} \in \mathbb{R}^{Q\times 1},\ \xi_{ijk}^{(\mathrm{W})} \in \mathbb{R}^{Q\times 1} \\ \Lambda_{jk}^{(\mathrm{B})} \in \mathbb{R}^{P\times Q},\ \Lambda_{ijk}^{(\mathrm{W})} \in \mathbb{R}^{P\times Q} \end{matrix}, \tag{3.17}$$

where Q denotes the total number of latent constructs and P denotes the total number of observed indicators. Further, configural invariance is also assumed across clusters as well as across latent classes.

2) Cross-level measurement invariance for factor loadings, such that the loadings are invariant at the within-level and the between-level for any given cluster:

$$\mathbf{\Lambda}_{k}^{(\mathrm{B})} = \mathbf{\Lambda}_{jk}^{(\mathrm{W})}. \tag{3.18}$$

The cross-level loading invariance is crucial to ensure that the between-level latent construct can be meaningfully interpreted as the configural construct and that it shares the same interpretation across clusters (Jak, 2019; Jak et al., 2013, 2021; Jak & Jorgensen, 2017).

3) Strong factorial (scalar) invariance across the clusters within each latent class, such that the factor loadings are invariant within the same latent class

$$\mathbf{\Lambda}_{jk} = \mathbf{\Lambda}_{j'k} = \mathbf{\Lambda}_{k}. \tag{3.19}$$

By doing so, it is ensured that the variance of the within-level latent construct is comparable across clusters, and thus the between-level random effects of the latent construct variability can be meaningfully interpreted. Further, the between-level residual

variance $\boldsymbol{\Theta}_k^{(B)}$ is constrained to be zero to ensure that within each class, the observed differences in cluster means are all explained by the between-level configural construct[5].

4) Weak factorial (metric) invariance among the latent classes, such that the factor loadings are invariant across the latent classes:

$$\boldsymbol{\Lambda}_k = \boldsymbol{\Lambda}_{k'} = \boldsymbol{\Lambda}. \tag{3.20}$$

This loading equality ensures that the fixed effects and random effects of the within-level latent construct variability is comparable across the latent classes. Theoretically, other more restrictive forms of measurement invariance cross the latent classes can also be imposed and tested, including strong factorial invariance and strict invariance. But they will not be further discussed in the current study, as metric invariance suffices for the purpose of variability modeling.

For simplicity and to reduce the number of parameters being estimated, the current study will further assume invariance in within-level residual variance $\Theta_k^{(W)}$ across the clusters within same latent class. That is, the observed heterogeneity in within-level variability is only driven by the randomly varying variability of the within-level focal construct, rather than other hidden factors that are not intended to be measured. But it is certainly not a necessary condition in order to meaningfully compare and interpret the variability of latent constructs. As a side note, some of the aforementioned invariance conditions may be overly restrictive and can potentially be further relaxed. For example, it is possible to specify a saturated between-level model without assuming

[5] This assumption may seem overly restrictive for the purpose of modeling variability per se. But with strong invariance across clusters, the model complexity is reduced and interpretability is greatly improved. The possibility of relaxing this assumption will be investigated in future research.

cross-level measurement invariance, if the between-level latent construct is not of theoretical interest.

3.4 Model Assumptions

The theoretical validity of MMSEM-RV and meaningful interpretation of the estimation results rely on a collection of important assumptions. The assumptions relevant for the current study are briefly summarized and discussed below.

1) Measurement invariance. As detailed in section 3.3, a series of measurement invariance assumption is critical to ensure the interpretability of the latent constructs and the relevant model parameters.

2) Correct model specification. It is assumed that the analytical model is correctly specified with respect to the measurement model as well as the structural model. That is, the factor structure and residual structure are correctly specified for the observed indicators. Additionally, when applicable, the within-level and between-level structural relations are correctly specified.

3) No omitted covariates. For the current study, it is assumed that there are no omitted covariates at any levels that can potentially lead to spurious clustering or artificial heteroskedasticity.

4) No level-1 latent classes. The latent sub-populations are only defined at the between level. There are no unobserved sub-populations within clusters.

5) Continuous observed measures. The observed variables are inherently continuous.

6) Multivariate normality. For the current study, it is assumed that all the residuals and latent random effects are continuously and multivariate normally distributed.

7) No missing data. For simplicity, the current study assumes that the observed data are complete without missingness.

Chapter 4. Simulation Studies

4.1 Simulation Design

The current study utilizes simulation to systematically evaluate the performance of MMSEM-RV with Bayesian estimation under varying manipulated conditions. The simulation studies consist of two parts: the first part is based on population analysis and the second part is based on Monte Carlo methods with repeated sampling. The detailed simulation design is summarized in this section. The simulation studies aim to address the following research questions:

1. Can MMSEM-RV be effectively estimated using a Bayesian approach via Gibbs sampling, yielding a converged solution?
2. Can the focal model parameters of MMSEM-RV be accurately and efficiently estimated?
3. How accurate is the latent class assignment when MMSEM-RV is used for classifying clusters, and how uncertain is the classification decision being made?
4. To what extent do factors such as sample size, class separation, and measurement quality impact the estimation of MMSEM-RV?
5. How does the proposed MMSEM-RV model compare to a naïve mean score based approach that ignore the measurement error?

4.1.1 Data Generating Model

For the simulation studies, data are all generated based on the following unconditional MMSEM-RV model with a fixed number of observed indicators ($P = 4$) that measure one common latent factor at both the within and between levels:

$$
\begin{aligned}
\mathbf{Y}_{ij}\Big|_{C_j=k} &= \boldsymbol{\alpha}_k + \boldsymbol{\Lambda}\xi_{jk}^{(\mathrm{B})} + \boldsymbol{\varepsilon}_{jk}^{(\mathrm{B})} + \boldsymbol{\Lambda}\xi_{ijk}^{(\mathrm{W})} + \boldsymbol{\varepsilon}_{ijk}^{(\mathrm{W})} \\
\boldsymbol{\varepsilon}_{jk}^{(\mathrm{B})} &\sim \mathrm{N}\,(0, \Theta_k^{(\mathrm{B})}) \\
\boldsymbol{\varepsilon}_{ijk}^{(\mathrm{W})} &\sim \mathrm{N}\,(0, \Theta_k^{(\mathrm{W})}) \\
\boldsymbol{\xi}_{jk}^{*(\mathrm{B})} &= \boldsymbol{\gamma}_k^{(\mathrm{B})} + \mathbf{v}_{jk}^{(\mathrm{B})} \\
&= \begin{bmatrix} \xi_{jk}^{(\mathrm{B})} \\ \ln(\sigma_{jk}^2) \end{bmatrix} = \begin{bmatrix} 0 \\ \gamma_{10k} \end{bmatrix} + \begin{bmatrix} \mathrm{v}_{0jk} \\ \mathrm{v}_{1jk} \end{bmatrix} \qquad . \qquad (4.1) \\
\xi_{ijk}^{(\mathrm{W})}\Big|_{\sigma_{jk}^2} &\sim \mathrm{N}\,(0, \sigma_{jk}^2) \\
\mathbf{v}_{jk}^{(\mathrm{B})} &= \begin{bmatrix} \mathrm{v}_{0jk} \\ \mathrm{v}_{1jk} \end{bmatrix} \sim \mathrm{N}\,(\mathbf{0}, \Phi_{\mathrm{v}k}) = \mathrm{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \phi_{0k} & \\ \phi_{10k} & \phi_{1k} \end{bmatrix}\right) \\
P(C_j = k) &= \pi_{jk} = \frac{\exp(\tau_k)}{\sum_k \exp(\tau_k)}
\end{aligned}
$$

The population value of the model parameters are specified to reflect what researchers can potentially encounter in realistic settings. In summary, the following parameters are set to a fixed value in the data generation process across the simulation conditions:

1) The first factor loading is fixed to one. Within the same condition, the other factor loadings are constrained to be equal across the rest of the observed indicators.
2) The between-level and within-level residual covariance is each fixed to be a diagonal matrix with homogeneous elements across the observed indicators. They are also constrained to be constant across the latent classes:

$$\Theta_k^{(B)} = \begin{bmatrix} 0 & & & \\ & 0 & & \\ & & 0 & \\ & & & 0 \end{bmatrix}, \Theta_k^{(W)} = \begin{bmatrix} 0.1 & & & \\ & 0.1 & & \\ & & 0.1 & \\ & & & 0.1 \end{bmatrix} \quad \forall k .$$

Importantly, assuming strong measurement invariance across clusters, the elements of between-level residual variance matrix $\Theta_k^{(B)}$ are all fixed to be zero.

3) The total number of latent classes is fixed to two across all the simulated conditions.

4) The between-level random effects covariance in latent class 1 and latent class 2 are respectively specified as

$$\Phi_{v1} = \begin{bmatrix} 0.3 & \\ -0.03 & 0.03 \end{bmatrix}, \Phi_{v2} = \begin{bmatrix} 0.5 & \\ -0.05 & 0.05 \end{bmatrix}.$$

As such, in the population there is less variation among the clusters in latent class 1 compared to the clusters in latent class 2, regarding both the between-level latent construct and the within-level variability.

5) The class size ratio between latent class 1 and latent class 2 is set to 2:3, corresponding to unconditional mixing proportions of $\pi_1 = 0.4$ and $\pi_2 = 0.6$.

6) The data generating model has invariant factor loadings cross-level and cross-class. For any cluster, the within-level and between-level loadings are both equal to a fixed factor loading: $\mathbf{\Lambda}^{(W)} = \mathbf{\Lambda}^{(B)} = \mathbf{\Lambda}$.

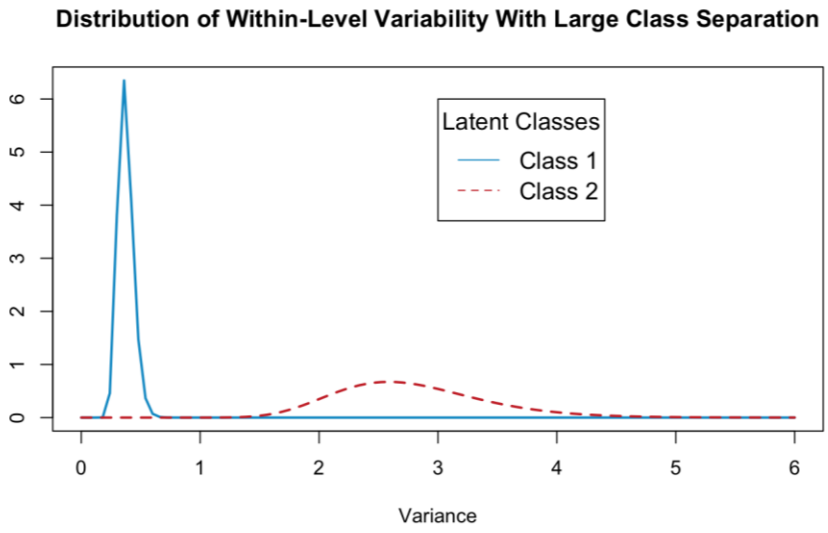
7) Class separation in the average within-level variability is fixed at $\gamma_{101} = -1, \gamma_{102} = 1$ (corresponding to an effect size ranging from 9.75 to 9.76, which varies with different sample size[6]). On average, clusters in latent class 1 are highly cohesive, while the

[6] While the magnitude of this effect size may seem extreme on Cohen's scale, it should be noted that this measure gauges the class separation in terms of log variance, rather than means. Therefore, Cohen's

within-cluster observations are much more heterogeneous for the clusters in latent class 2. Figure 4-1 illustrates the distribution of the within-level variance σ^2_{jk} given the designated class separation.

Figure 4-1

Distribution of the Within-Level Variance in the Simulation Study



4.1.2 Manipulated Factors

Given the population model in Eq. (4.1) and the aforementioned fixed parameter values, the following factors are systematically manipulated to form different simulated conditions under a full factorial design:

conventional classification of effect sizes may not be directly applicable here. The current study only aims to investigate the impact of intercept separation on model estimation, conditional on large class separation in terms of latent random variability. Future work can further examine the impact of varying degrees of variability separation. It is expected, though, that as the classes get more similar in the distribution of random variability, it will become increasingly challenging to estimate MMSEM-RV. Yet, the extent to which the model remains robust at different levels of variability separation remains an open area warranting further investigation.

1) The study varies the total number of clusters across four levels: $J = 60$, 100, 300, or 600. In the literature of mixture MSEM, the number of clusters are usually varied between 20 and 100 (e.g., Kim & Wang, 2019), while the empirical data example of PISA has 162 clusters. Considering the structural complexity of MMSEM, the model performance is assessed with larger number of clusters in the current study than those in previous research.
2) The cluster size is manipulated with two levels: $n_j = 20$ or 80. The actual number of observations within each cluster is then randomly sampled from a discrete uniform distribution with the support of $(n_j - 5, n_j + 5)$. The cluster size of 20 resembles the statistics of PISA U.S. data, while larger values have been specified in prior simulations studies on mixture MSEM (see De Roover et al., 2020).
3) Latent construct reliability is manipulated by varying the factor loadings. The unstandardized factor loadings are varied with two levels, $\lambda_i = 0.3$ or 0.8 ($i = 2, 3, 4$), indicating different levels of construct reliability (see more details below).
4) The grand intercepts in each latent class are varied with two levels: (a) large intercept separation $\boldsymbol{\alpha}_1 = (8,\ 8,\ 8,\ 8)'$, $\boldsymbol{\alpha}_2 = (0,\ 0,\ 0,\ 0)'$, and (b) moderate intercept separation $\boldsymbol{\alpha}_1 = (2,\ 2,\ 2,\ 2)'$, $\boldsymbol{\alpha}_2 = (0,\ 0,\ 0,\ 0)'$. On average, the observed indicators have higher population means in latent class 1 than those in latent class 2.

In summary, a total number of 32 manipulated conditions are considered for the simulation study (Table 4-1).

Table 4-1

Manipulated Factors and Levels in Simulation Study

Manipulated Factors	Levels
Number of clusters	$J = 60$, 100, 300, or 600
Cluster size	$n_j = 20$ or 80
Intercept separation	$\alpha_1 = 2$ or 8; $\alpha_2 = 0$
Measurement quality	Unstandardized factor loadings $\lambda = 0.3$ or 0.8

Although latent construct reliability and intraclass correlation coefficient (ICC) are not purposefully manipulated in the current study, they can still vary across the simulated conditions as they are functions of the manipulated factors. It is worth noting that, as MMSEM-RV allows the within-level variability to randomly vary across clusters, the within-level latent construct reliability and ICC are thus both random variables rather than a fixed value for each simulated condition. To quantify the reliability in latent construct and measured variables, the current study defines and reports the "average" within-level construct reliability for latent class *k* using coefficient omega (i.e., composite reliability)

$$\omega_k^{(W)} = \frac{(\sum_{i=1}^{P} \lambda_{ik}^{(W)})^2 \exp(\gamma_{10k})}{(\sum_{i=1}^{P} \lambda_{ik}^{(W)})^2 \phi_{0k} + \sum_{i=1}^{P} \boldsymbol{\Theta}_{iik}^{(W)}} \quad , \tag{4.2}$$

and the "average" ICC of the p^{th} observed indicator

$$\text{ICC}_{pk} = \frac{(\lambda_{ik}^{(B)})^2 \phi_{0k} + \boldsymbol{\Theta}_{iik}^{(B)}}{(\lambda_{ik}^{(B)})^2 \phi_{0k} + \boldsymbol{\Theta}_{iik}^{(B)} + (\lambda_{ik}^{(W)})^2 \exp(\gamma_{10k}) + \boldsymbol{\Theta}_{iik}^{(W)}} \quad , \tag{4.3}$$

both formulated as a function of the mean within-level variability[7]. Table 4-2 summarizes the values of "average" within-level composite reliability and "average" ICC under different simulation conditions. Given constant within-level residual variance, higher factor loading is

[7] Despite the name, they are, however, not the actual average of these random variables. As Jensen's inequality suggests, for a convex function g we have $E[g(X)] \geq g(E[x])$. Therefore, the expectations of the reliability and ICC are not equal to the values obtained by plugging in the fixed effects of within-level variability (hence they are referred to as the *"average"* reliability and *"average"* ICC with quotation marks).

expected to yield better within-level construct reliability for the same latent class; this contrast is particularly salient for latent class 1, which is designated to have smaller within-level construct variance. Similarly, with constant within-level residual variance, the "average" ICC of observed indicators should also increase as the factor loading goes up. For reference, previous simulation studies in the context of mixture MSEM usually have an ICC value ranging from 0.05 to 0.25, which are the common ICC values found in national education studies (Hedges & Hedberg, 2007; Kim & Wang, 2019).

At the between-level, because strong measurement invariance is assumed, the between-level construct reliability

$$\omega_k^{(\mathrm{B})} = \frac{(\sum_{i=1}^{P} \lambda_{ik}^{(\mathrm{B})})^2 \phi_{0k}}{(\sum_{i=1}^{P} \lambda_{ik}^{(\mathrm{B})})^2 \phi_{0k} + \sum_{i=1}^{P} \boldsymbol{\Theta}_{iik}^{(\mathrm{B})}} \tag{4.4}$$

remains constant across the simulation conditions. More specifically, the between-level construct has perfect construct reliability $\omega_k^{(\mathrm{B})} = 1$ in all conditions as the between-level residual variance is fixed to zero.

Table 4-2

The Value of Composite Reliability and ICC under Different Conditions

Unstandardized Factor Loading	"Average" Within-level Reliability		"Average" ICC for Non-scaling Indicators	
	Class 1	Class 2	Class 1	Class 2
$\lambda_i = 0.3$	$\omega_1^{(\mathrm{W})} = 0.769$	$\omega_2^{(\mathrm{W})} = 0.961$	$\mathrm{ICC}_{p1} = 0.169$*	$\mathrm{ICC}_{p2} = 0.115$
$\lambda_i = 0.8$	$\omega_1^{(\mathrm{W})} = 0.914$	$\omega_2^{(\mathrm{W})} = 0.987$	$\mathrm{ICC}_{p1} = 0.364$	$\mathrm{ICC}_{p2} = 0.148$

Note. **p* is the index for observed indicator, *p* = 2, 3, or 4. The between-level reliability $\omega_k^{(\mathrm{B})}$ is equal to 1 across all the simulated conditions as strong measurement invariance is assumed. The "average" ICC for the scaling indicator is constant across all the simulated conditions as the factor loading is always constrained to be 1: for the scaling indicator, the "average" ICC is 0.39

in latent class 1 and 0.15 in latent class 2, regardless of the manipulated factor loading conditions.

4.1.3 Analytical Model and Estimation

The unconditional MMSEM-RV specified in Eq. (4.1) is fit to the simulated data with Bayesian estimation, implemented using Gibbs sampling in `rjags` (Plummer et al., 2022). The number of latent classes (k) is set to a known value of two, assuming a decision informed by relevant prior knowledge. To reduce the risk of label switching, the model is estimated with one single chain. The estimation process terminates after 50,000 iterations post the burn-in period. For Bayesian estimation, the following priors are employed in the current study:

1) Diffuse prior is used for the fixed effects of log within-level variance: $\gamma_{101} \sim N(-1, 100000)$ and $\gamma_{102} \sim N(1, 100000)$. The identifiability constraints are imposed on the fixed effects $\gamma_{101} < ... < \gamma_{10K}$ to prevent label switching from iteration to iteration.
2) Diffuse prior is used for the free factor loadings: $\lambda_i \sim N(0, 10000000)$, $i = 2, 3, 4$.
3) Informative normal prior is used for item intercepts: $\alpha_1 \sim N(8, 10)$ or $\alpha_1 \sim N(2, 10)$, and $\alpha_2 \sim N(0, 10)$.
4) The inverse Wishart distribution is used for between-level random effects variance-covariance matrix: $\boldsymbol{\Phi}_{vk} \sim W^{-1}(\mathbf{I}_2, 6)$.
5) Non-informative symmetric Dirichlet prior is used for the mixing proportion: $\boldsymbol{\pi} \sim \text{Dir}(\boldsymbol{\alpha} = (1, 1))$.
6) Informative inverse gamma prior is used for within-level residual variance: $\Theta^{(W)} \sim \text{Inv-Gamma}(3, 0.5)$.

The model convergence is assessed primarily using Geweke's diagnostic (Geweke, 1992), followed by visual inspection of the trace plots. Each parameter's convergence is assessed

separately based on the marginal posterior distribution; the model is deemed converged only if the estimation for all the model parameters reaches convergence. Given the large number of parameters involved and the corresponding multiple hypothesis testing in Geweke's diagnostic, Bonferroni correction is applied to control for family wise error rate. When the model converges, the posterior distributions for each model parameter is summarized and further examined. More specifically, the posterior means are reported as the parameter point estimates and the 95% highest posterior density intervals (HPDI) are used as the interval estimates. For classification purpose, the posterior mode of the latent class assignment for each cluster *j* is used as the predicted class membership $\hat{C}_j$.

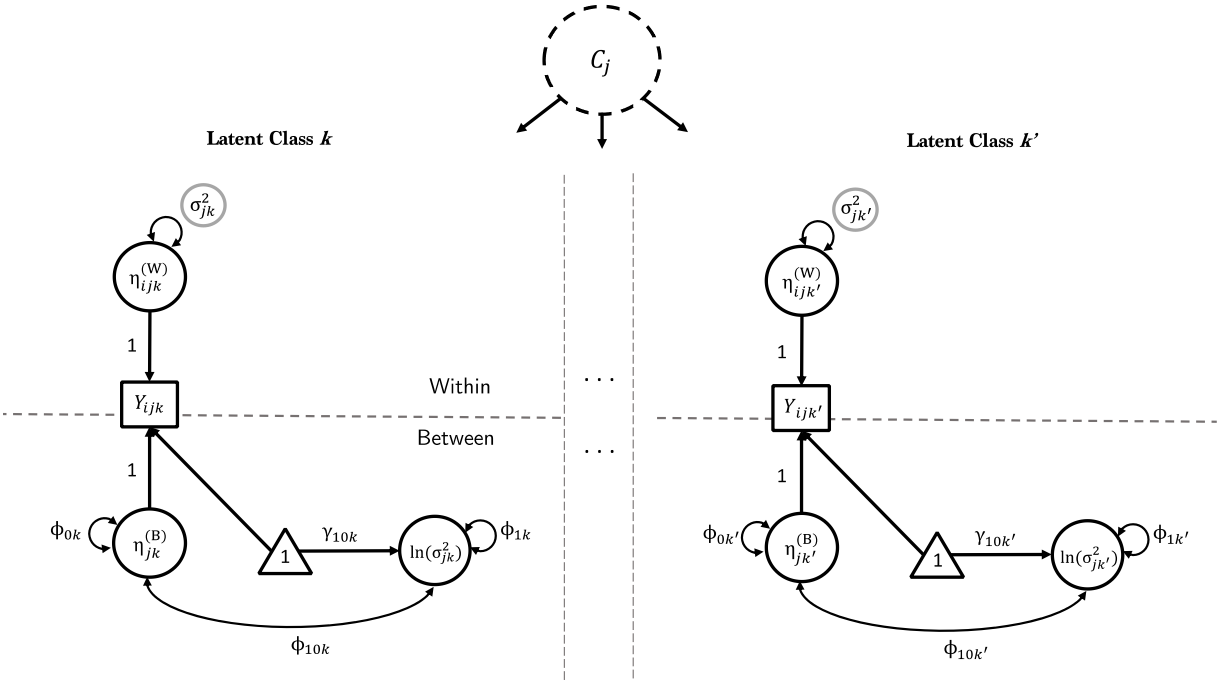
In practice, due to various reasons, it is not uncommon for researchers to compute a composite score across the observed indicators rather than explicitly modeling their measurement structure. The naïve approach with composite score is practically appealing as it simplfies the model specification and estimation. However, it assumes the observed indicators are measured without error, and they perfectly reflect the underlying latent construct. To assess the potential consequences of ignoring measurement error, a naïve mean score based MMSEM-RV is also fit to the simulated data:

$$
\begin{aligned}
Y_{ij}\Big|_{C_j=k} &= \alpha_k + \eta_{jk}^{(\mathrm{B})} + \eta_{ijk}^{(\mathrm{W})} \\
\eta_{ijk}^{(\mathrm{W})}\Big|_{\sigma_{jk}^2} &\sim \mathrm{N}\,(0, \sigma_{jk}^2) \\
\eta^{*(\mathrm{B})}_{jk} &= \boldsymbol{\gamma}_k^{(\mathrm{B})} + \mathbf{v}_{jk}^{(\mathrm{B})} \\
&= \begin{bmatrix} \eta_{jk}^{(\mathrm{B})} \\ \ln(\sigma_{jk}^2) \end{bmatrix} = \begin{bmatrix} 0 \\ \gamma_{10k} \end{bmatrix} + \begin{bmatrix} \nu_{0jk} \\ \nu_{1jk} \end{bmatrix} \\
\mathbf{v}_{jk}^{(\mathrm{B})} &= \begin{bmatrix} \nu_{0jk} \\ \nu_{1jk} \end{bmatrix} \sim \mathrm{N}\,(\mathbf{0}, \Phi_{\nu k}) = \mathrm{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \phi_{0k} & \\ \phi_{10k} & \phi_{1k} \end{bmatrix} \right) \\
P(C_j = k) &= \pi_{jk} = \frac{\exp(\tau_k)}{\sum_k \exp(\tau_k)}
\end{aligned}
\quad , \tag{4.5}
$$

where a composite score Y_{ij} is computed as the arithmetic mean of the observed indicators Y_1 through Y_4. With this model, Y_{ij} is decomposed into the grand mean α_k, between-level cluster deviation from the grand mean ($\eta_{jk}^{(\mathrm{B})}$), and within-level unit deviation from the cluster mean ($\eta_{ijk}^{(\mathrm{W})}$), all of which are treated as if they were measured without error. The within-level variability is thus modeled as the log transformed variance of the within-level deviation $\eta_{ijk}^{(\mathrm{W})}$. Therefore, if $\eta_{ijk}^{(\mathrm{W})}$ contains any measurement error, it is expected to inflate the variability estimates. For illustration, the path diagram for this mean score model is provided in Figure 4-2. The mean score model is also estimated using Bayesian approach, following the similar procedures and prior specifications as outlined above. The only difference is that fewer parameters are involved given the simplified model. For example, there is no need to specify the priors for factor loadings and within-level residual variance-covariances for the mean score model. The convergence diagnostic and results summary are also carried out in a similar manner as previously described.

Figure 4-2

Path Diagram for Unconditional Mean Score MMSEM-RV



4.2 Simulation Study Part I: Population Analysis

In the first part of the simulation study (part I), population analysis was conducted for all the 32 conditions under the full factorial design. To implement population analysis, a simulated dataset was first randomly generated based on the population data generating model. Subsequently, Bollen-Stine transformation (Bollen & Stine, 1992) was employed to ensure the sample statistics of the observed data precisely match the designated population values. This procedure is details below.

For a general definition, with mean-centered variables **X** let us denote the sample covariance matrix of **X** by **S,** which is expected to deviate from the population covariance matrix **Σ** due to sampling error with a limited sample size. The Bollen-Stine transformation

$$\mathbf{Z} = \mathbf{X}\mathbf{S}^{-1/2}\mathbf{\Sigma}^{1/2} \tag{4.6}$$

can be used to obtain a new set of variables $\mathbf{Z}$ that has a covariance matrix equal to the population value $\mathbf{\Sigma}$. For the specific purpose of population analysis within the current study, the following steps were followed to generate the analytical datasets:

Step 1. Generation of between-level random data. In the first step, data for the latent variables $\mathbf{v}_{jk}^{(\mathrm{B})}$, $\boldsymbol{\varepsilon}_{jk}^{(\mathrm{B})}$, and $\boldsymbol{\varepsilon}_{ijk}^{(\mathrm{W})}$ were randomly generated from multivariate normal distributions based on the population parameter values specific to each simulated condition.

Step 2. Transformation of between-level random effects. In this step, within each latent class, the randomly generated random effects $\mathbf{v}_{jk}^{(\mathrm{B})}$ were mean centered and transformed using Eq. (4.6), with results denoted as $\tilde{\mathbf{v}}_{jk}^{(\mathrm{B})}$. This step ensures the strict equivalence between the observed covariance matrix of $\tilde{\mathbf{v}}_{jk}^{(\mathrm{B})}$ and its population value Φ_{vk}. The between-level latent variables $\boldsymbol{\xi}_{jk}^{*(\mathrm{B})}$ were subsequently generated by combining the fixed effects $\boldsymbol{\gamma}_{k}^{(\mathrm{B})}$ and the transformed random effects $\tilde{\mathbf{v}}_{jk}^{(\mathrm{B})}$. As such, it is guaranteed that the covariance matrix of the between-level latent variables $\boldsymbol{\xi}_{jk}^{*(\mathrm{B})}$ is exactly equal to Φ_{vk}.

Step 3. Transformation of within-level residuals. At this step, the within-level residuals $\boldsymbol{\varepsilon}_{ijk}^{(\mathrm{W})}$ were mean centered and transformed according to Eq. (4.6), resulting in a new set of transformaed residuals $\tilde{\boldsymbol{\varepsilon}}_{ijk}^{(\mathrm{W})}$. This step ensures that the observed covariance matrix of $\tilde{\boldsymbol{\varepsilon}}_{ijk}^{(\mathrm{W})}$ is strictly equal to the population value $\Theta_{k}^{(\mathrm{W})}$.

Step 4. Transformation of within-level latent variables. In this step, at first the within-level latent construct $\xi_{ijk}^{(\mathrm{W})}$ were randomly generated from a normal distribution for each individual within each cluster. The variance of this normal distribution was determined by the

random variability latent variable $\ln(\sigma_{jk}^2)$ obtained in step 2. Subsequently, within each cluster, the within-level latent variables $\xi_{ijk}^{(W)}$ were transformed to ensure that the variance of the within-level latent constructs is strictly σ_{jk}^2.

Step 5. Generation of observed data. Finally, observed variables $\mathbf{Y}_{ij}$ were generated based on the data generating model specified in Eq. (4.1). Up to this point, all the latent variables have already been transformed to match their population distribution.

4.2.1 Outcome Evaluation

For each manipulated condition, both the unconditional latent construct MMSEM-RV specified in Eq. (4.1) and the naïve mean score based MMSEM-RV specified in Eq. (4.5) were fit to the simulated data. The results were summarized and evaluated with regard to the following aspects:

1) Model convergence—a dummy indicator of whether or not the model estimation reaches convergence.
2) Classification accuracy—the rate of correct class assignment, defined as:

$$\frac{\sum_{j=1}^{J} I(C_j = \hat{C}_j)}{J}, \tag{4.7}$$

where $\hat{C}_j$ is the estimated class membership for cluster j.

3) Classification certainty—the mean classification certainty (CC_{mean}) and minimum classification certainty ($CC_{\min}$), which are defined as the following respectively:

$$CC_{\text{mean}} = \frac{\sum_{j=1}^{J}\sum_{k=1}^{K}\hat{C}_{jk}\tilde{P}(C_{jk})}{J} \quad , \qquad (4.8)$$

$$CC_{\min} = \min_{j}\sum_{k=1}^{K}\hat{C}_{jk}\tilde{P}(C_{jk})$$

where $\hat{C}_{jk}$ is the dummy indicator denoting whether cluster *j* is predicted to be in latent class *k*, and $\tilde{P}(C_{jk})$ is the posterior classification probability of cluster *j* belonging to latent class *k*. This classification probability is computed as the posterior proportion of times when cluster *j* is assigned to latent class *k*.

4) Parameter estimation accuracy—the relative estimation bias for parameter θ, defined as:

$$\frac{\hat{\theta}-\theta}{|\theta|}. \qquad (4.9)$$

5) HPDI coverage—a dummy indicator denoting whether or not the 95% HPDI captures the population value of parameter θ:

$$I\left[\text{HPDI}_{\text{LL}}(\hat{\theta}) \le \theta \le \text{HPDI}_{\text{UL}}(\hat{\theta})\right]. \qquad (4.10)$$

4.2.2 Results

The population simulation analysis was implemented using R via the University of Maryland's HPC cluster Zaratan (a total number of 64 cores were requested for the population analysis jobs). The computing time to fit one unconditional latent construct MMSEM-RV model varies among the different conditions, ranging from as fast as 10.08 minutes with small sample size to as slow as 53.20 hours with large sample size. Comparatively, the naïve mean score based model is less computational intensive and thus took less time for model estimation, which ranges from 1.86 minutes to 1.34 hours across the conditions.

4.2.2.1 Latent Construct MMSEM-RV

According to the test results of Geweke's convergence diagnostic, the estimation of latent construct MMSEM-RV reached convergence in 27 out of 32 conditions (see Table B1 in Appendix B). The conditions that did not converge are not considered for outcome evaluation. For the model estimation that had converged, it shows satisfactory classification accuracy across all the conditions, ranging from 0.94 to 1 with perfect latent class prediction obtained in 20 conditions (see Table B1 in Appendix B). Overall, the classification decision is made with high level of certainty—the values of CC_{mean} and CC_{min} are both equal to 1 in all the converged conditions.

When MMSEM-RV is employed, for the specific purpose of modeling population heterogeneity in random variability, parameters that are typically of focal interest would include between-level fixed effects $\boldsymbol{\gamma}_k^{(\text{B})}$, between-level random effects covariance $\boldsymbol{\Phi}_{vk}$, and mixing proportions $\boldsymbol{\pi}$. The relative estimation bias for each of the focal parameters is presented in Figure Figure 4-4 to Figure 4-7. The fixed effects (γ_{101} and γ_{102}) and mixing proportions (π_1) are estimated with fairly good accuracy, yielding small relative bias with magnitude ranging from .0002 to .212. On the other hand, the variance and covariance of the between-level random effects are estimated with relatively larger bias, particularly for the variance of the random within-level variability, where the relative bias can get as high as 5.57. More specifically, the estimation tends to be more biased for this parameter in latent class 2, which was generated with large within-level variability.

Although the estimation bias varies from parameter to parameter, a common observation is that a large sample size can effectively reduce the estimation bias. Pooling across other manipulated factors, conditions with $J = 60$ or $J = 100$ yields larger average relative bias than

conditions with $J = 300$ or $J = 600$ for all the focal parameters. In terms of cluster size alone, its impact varies—while the conditions of $n_j = 80$ reduces estimation bias for certain model parameters (e.g., the between-level random effect variance), it does not help with some other model parameters. However, under the combined large sample size condition ($J = 600$ and $n_j = 80$), the relative bias is minimal, ranging from 0.0008 (for the mixing proportion $\boldsymbol{\pi}$) to 0.27 (for the variance of variability ϕ_{1k}).

Besides sample size, intercept separation also shows an impact on estimation accuracy. On average, large intercept separation ($\alpha_1 = 8$) generates smaller estimation bias for most model parameters (with the only exception being ϕ_{1k}). With regard to measurement quality, pooling across other manipulated factors, higher factor loading ($\lambda = 0.8$) in the measurement model leads to higher average relative bias across all the model parameters except for ϕ_{1k}. Such a negative effect of factor loading is particularly manifested when the item intercept separation is small ($\alpha_1 = 2$). When item intercepts are separated by 2, the model yields notably greater bias in parameter estimates when population factor loading λ is 0.8 compared to when it is 0.3. This interaction effect is further impacted by sample size—the discrepancy in relative bias between the condition of $\alpha_1 = 2$, $\lambda = 0.8$ and the condition of $\alpha_1 = 2$, $\lambda = 0.3$ diminishes when $J = 600$ and $n_j = 80$.

Figure 4-3

Relative Bias for Estimating the Probability of Belonging to Class 1

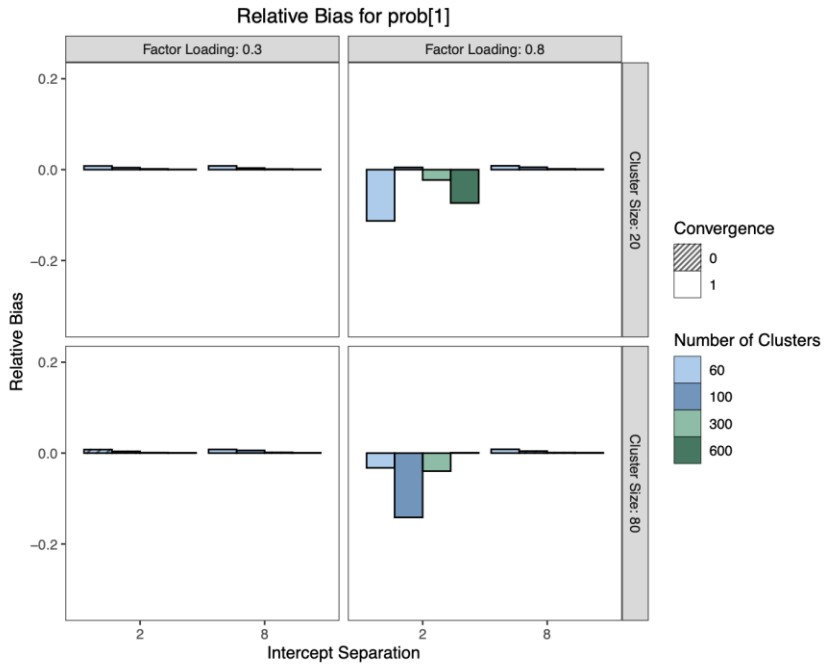
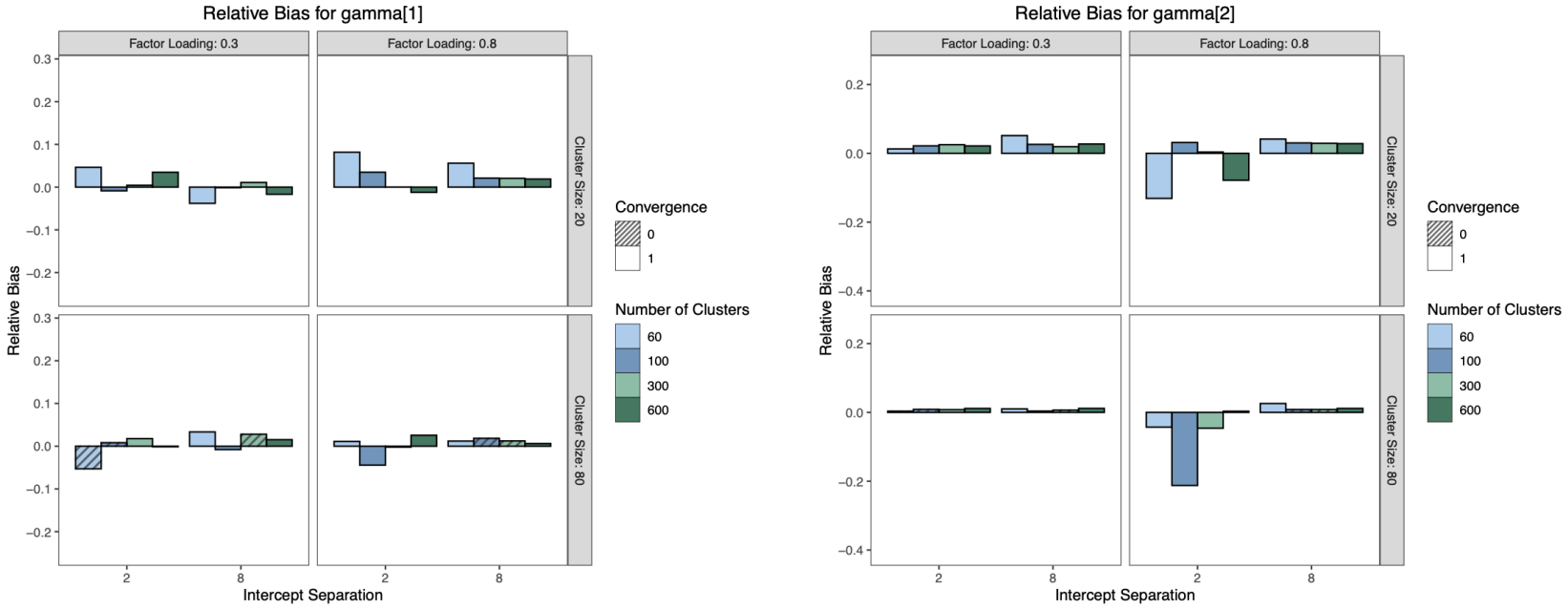


Figure 4-4

Relative Bias for Estimating Average Within-Level Log Variance in Class 1 and Class 2

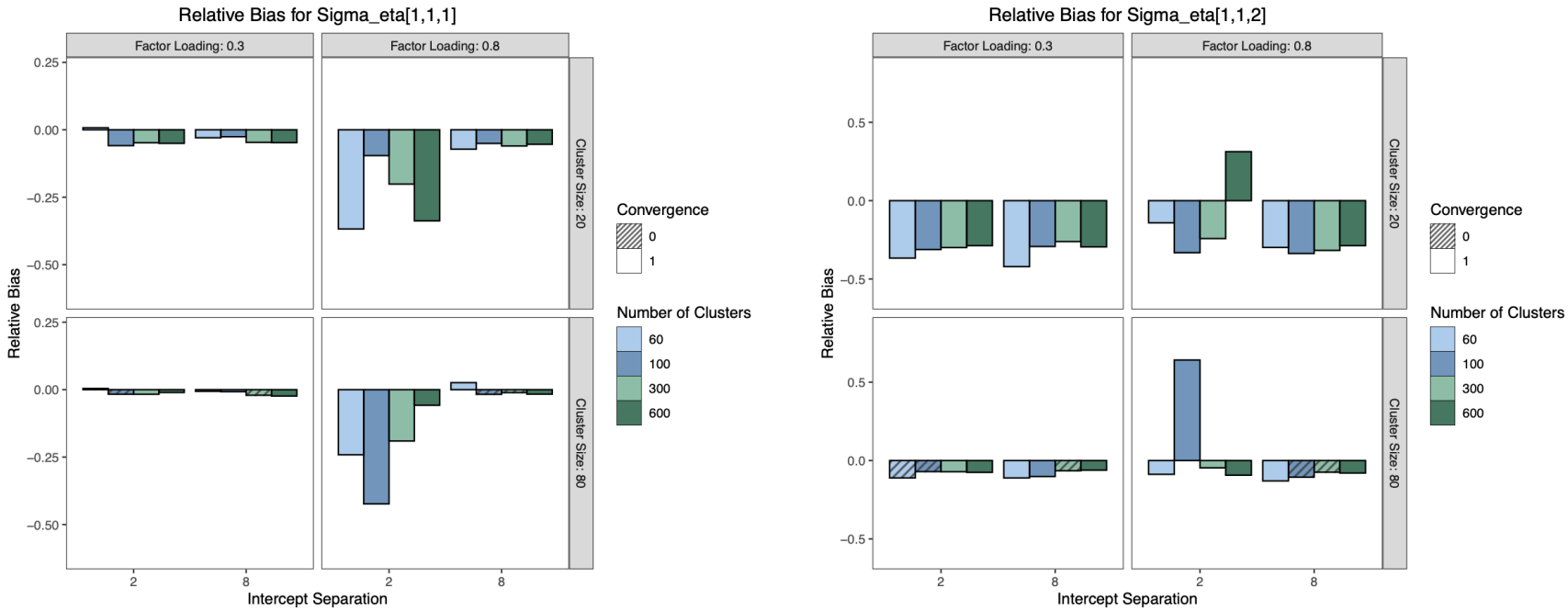


(a) (b)

Note. (a) Relative bias for estimating γ_{101}. (b) Relative bias for estimating γ_{102}.

Figure 4-5

Relative Bias for Estimating the Between-Level Latent Construct Variance in Class 1 and Class 2

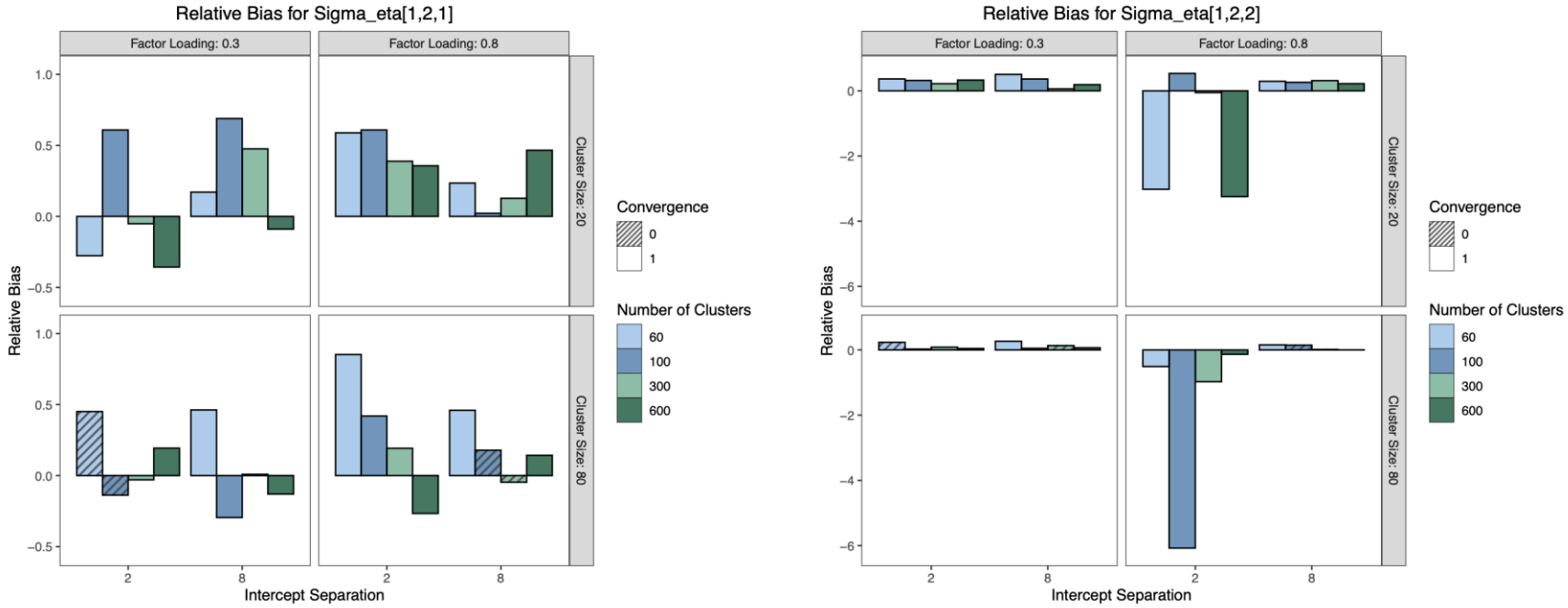


(a) (b)

Note. (a) Relative bias for estimating ϕ_{01}. (b) Relative bias for estimating ϕ_{02}.

Figure 4-6

Relative Bias for Estimating the Between-Level Random Effects Covariance in Class 1 and Class 2

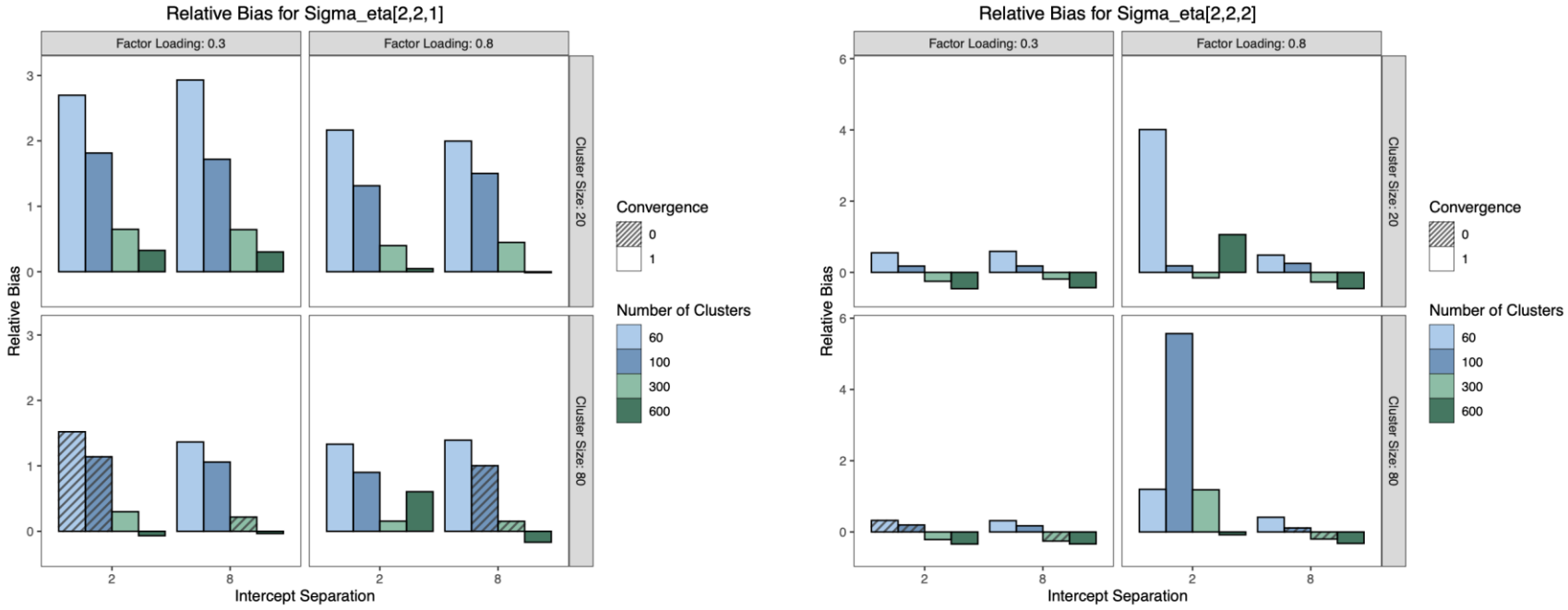


(a) (b)

Note. (a) Relative bias for estimating ϕ_{101}. (b) Relative bias for estimating ϕ_{102}.

Figure 4-7

Relative Bias for Estimating Variance of the Within-Level Log Variance in Class 1 and Class 2



(a) (b)

Note. (a) Relative bias for estimating ϕ_{11}. (b) Relative bias for estimating ϕ_{12}.

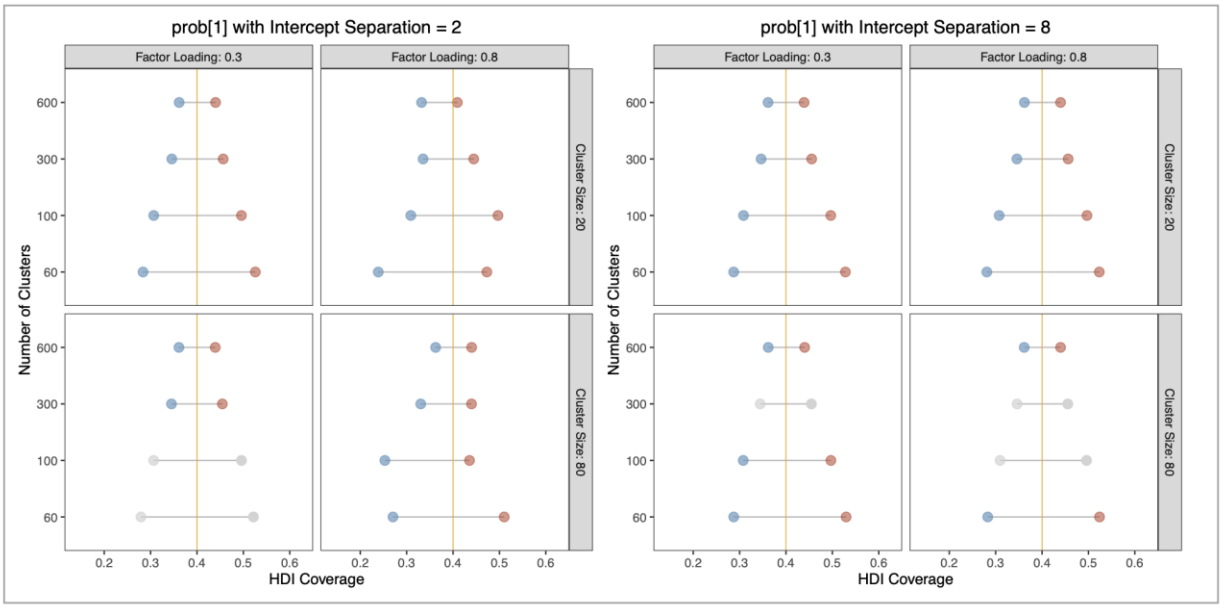
The 95% HPDIs for estimating each of the focal parameters under the population analysis are plotted in Figure 4-9 through Figure 4-11. As expected, given everything else equal, a larger sample size always leads to narrower interval width of the 95% HPDI. This effect is particularly noticeable when the number of clusters increases. The factor loading and intercept separation alone do not seem to impact the interval width of HPDI. With regard to the HPDI coverage, the 95% HPDI captures the population value of the mixing proportion π across all the conditions (Figure 4-8). The interval coverage is overall satisfactory for the fixed effects γ_{101} and γ_{102}, except in cases when the item intercept is separated by 2 and the factor loading is 0.8 (Figure 4-9). For between-level latent construct variance (Figure 4-10), the 95% HPDI also tend to fail to capture the population value in conditions where $\alpha_1 = 2$ and $\lambda = 0.8$; additionally, the interval coverage appears to be more problematic when the cluster size is small ($n_j = 20$) while the number of clusters is large ($J = 300$ or 600). Similarly, the interval coverage for the between-level covariance is generally good except for the case when the intercept separation is 2 while the factor loading is 0.8 (Figure 4-11). With regard to the variance of the within-level variability, the interval coverage is less satisfactory across the various conditions (Figure 4-12). For this variance parameter in latent class 1, the 95% HPDI only covers the population true value when there is a large number of clusters ($J = 300$ or 600); for the counterpart in latent class 2, the interval coverage is more problematic in cases where $\alpha_1 = 2$ and $\lambda = 0.8$, or when the number of clusters J is 600.

Although they are not of inferential interest, the relative bias and 95% HPDI coverage for estimating the other peripheral model parameters are provided in Figures B1-B10 in Appendix B. Bias is in general minimal for the nuisance parameters, ranging from .00002 to .280 across all the conditions. Comparatively speaking, the residual variances are estimated with slightly larger

bias than the other peripheral parameters. The 95% HPDI covers the population true values for the peripheral parameters in most of the simulated conditions. The interval coverage performance is comparatively less satisfactory regarding the factor loadings when the population is generated with $\alpha_1 = 2$ and $\lambda = 0.8$.

Figure 4-8

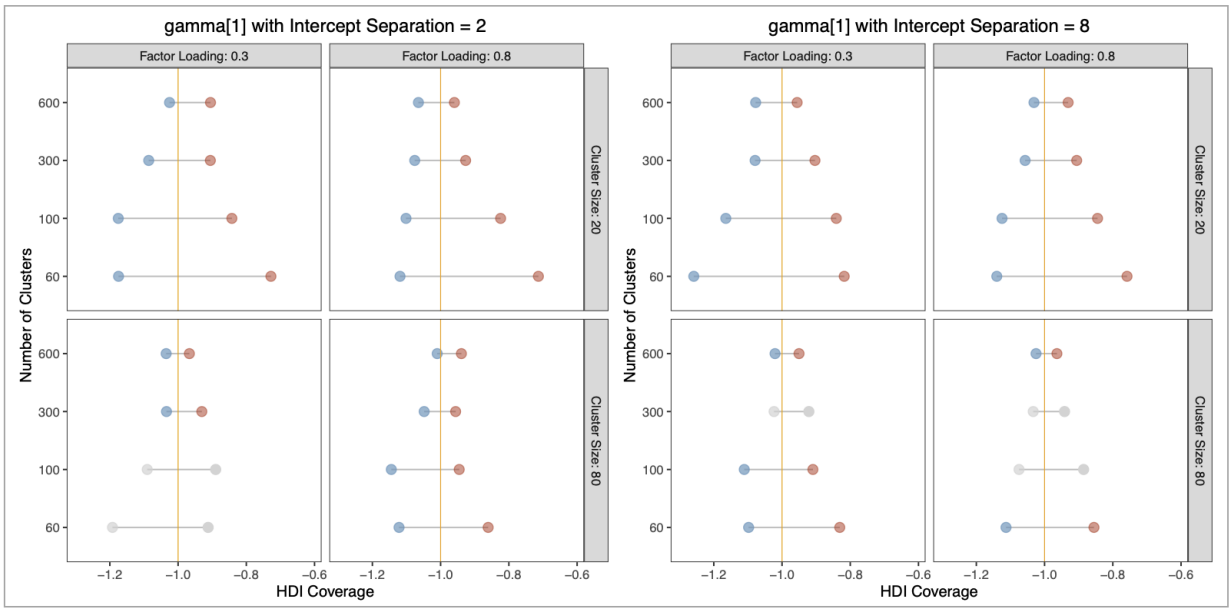
95% HPDI Coverage for Estimating the Probability of Belonging to Class 1



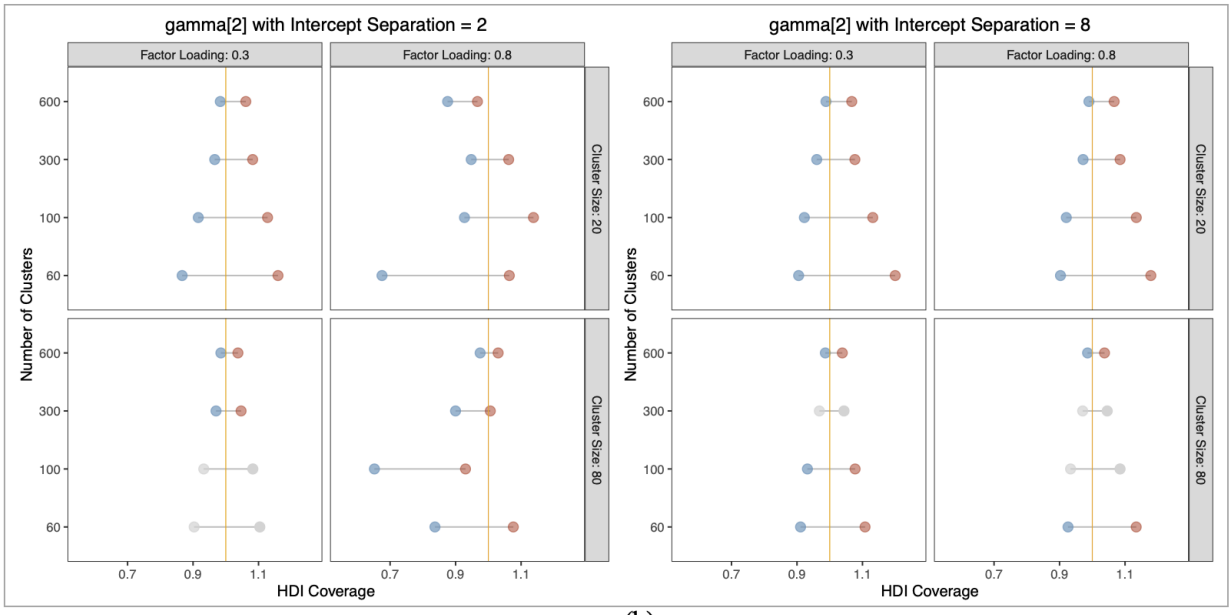
Note. Conditions that did not reach convergence are marked in light grey. Population value of the parameter is indicated by the golden line.

Figure 4-9

95% HPDI Coverage for Estimating Average Within-Level Log Variance in Class 1 and Class 2



(a)

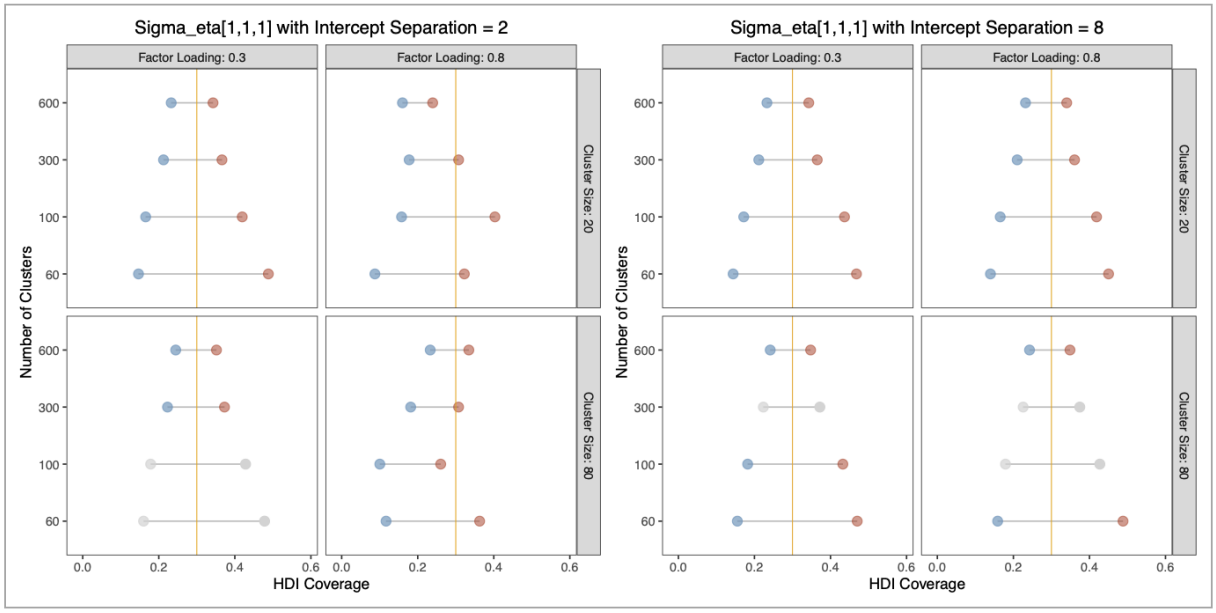


(b)

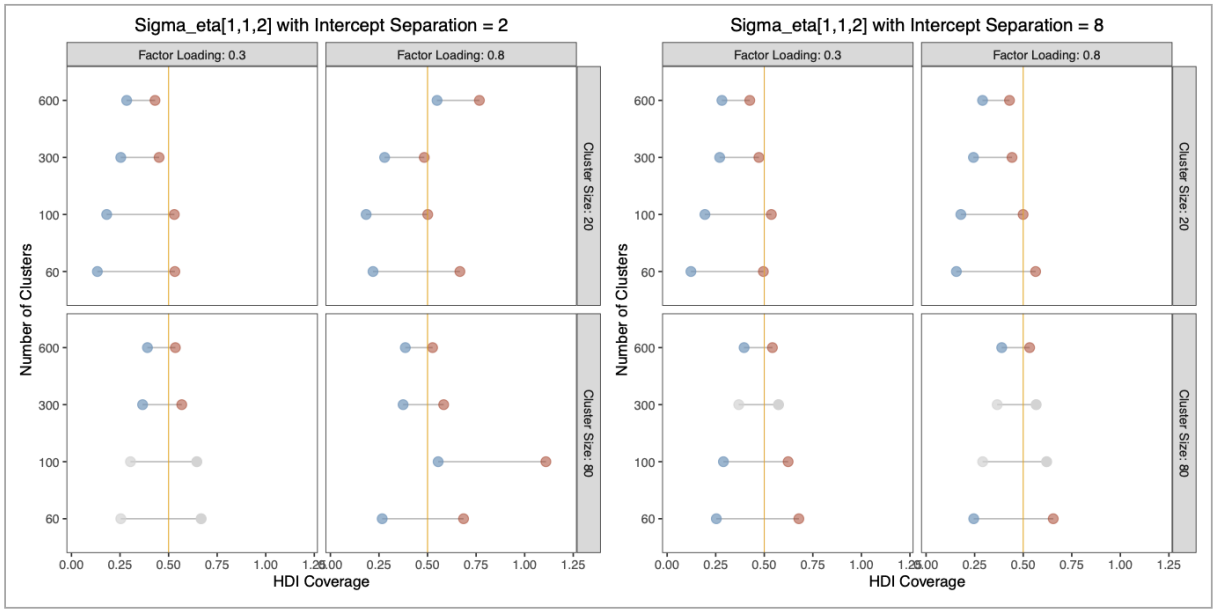
Note. (a) HPDI for estimating γ_{101}. (b) HPDI for estimating γ_{102}. Conditions that did not reach convergence are marked in light grey. Population value of the parameter is indicated by the golden line.

Figure 4-10

95% HPDI Coverage for Estimating the Between-Level Latent Construct Variance in Class 1 and Class 2



(a)

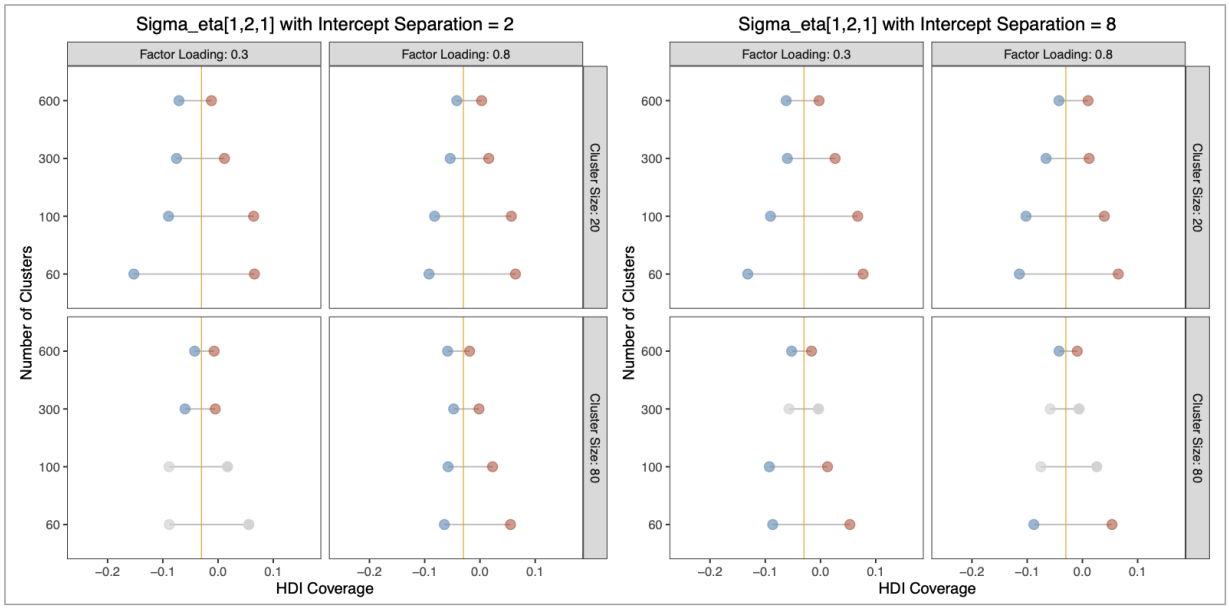


(b)

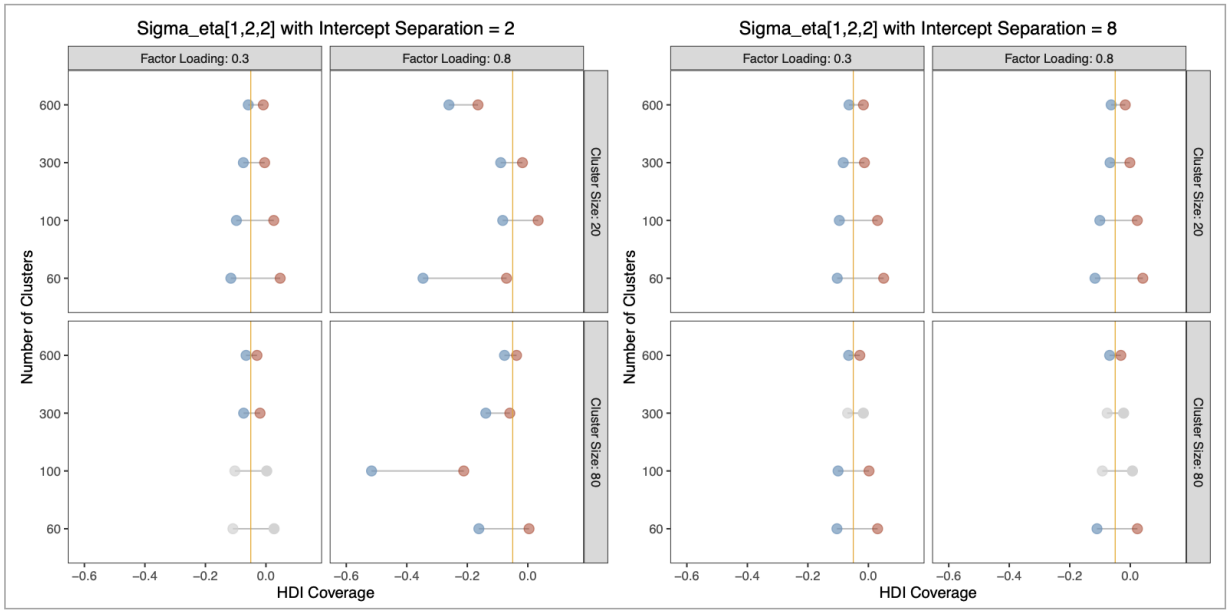
Note. (a) HPDI for estimating ϕ_{01}. (b) HPDI for estimating ϕ_{02}. Conditions that did not reach convergence are marked in light grey. Population value of the parameter is indicated by the golden line.

Figure 4-11

95% HPDI Coverage for Estimating the Between-Level Random Effects Covariance in Class 1 and Class 2



(a)

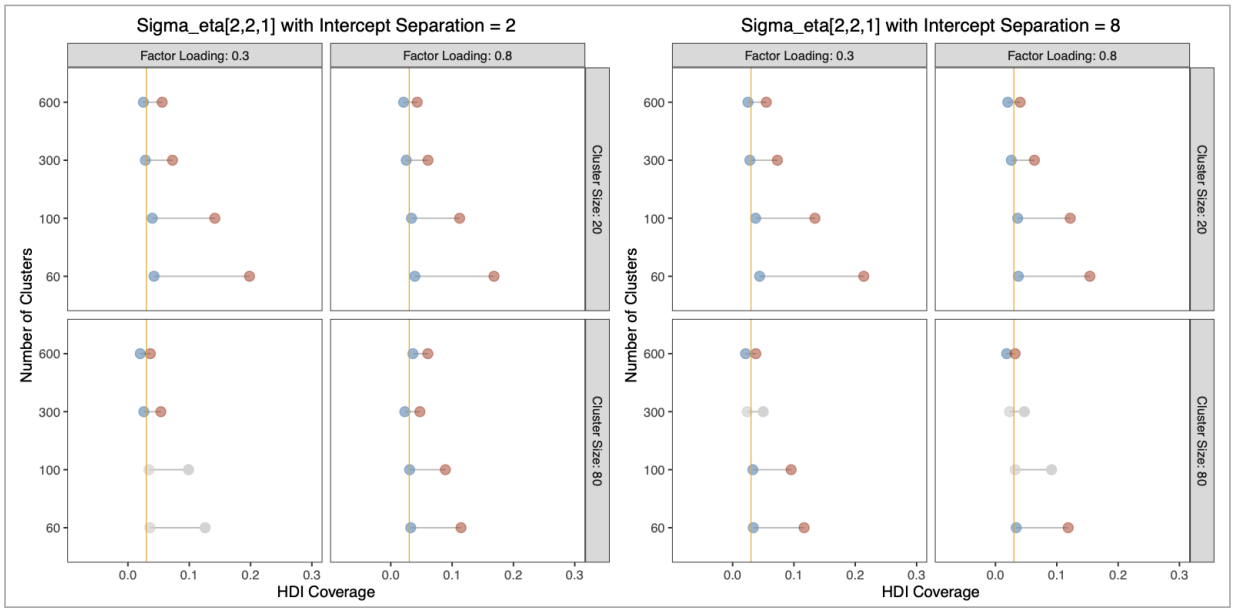


(b)

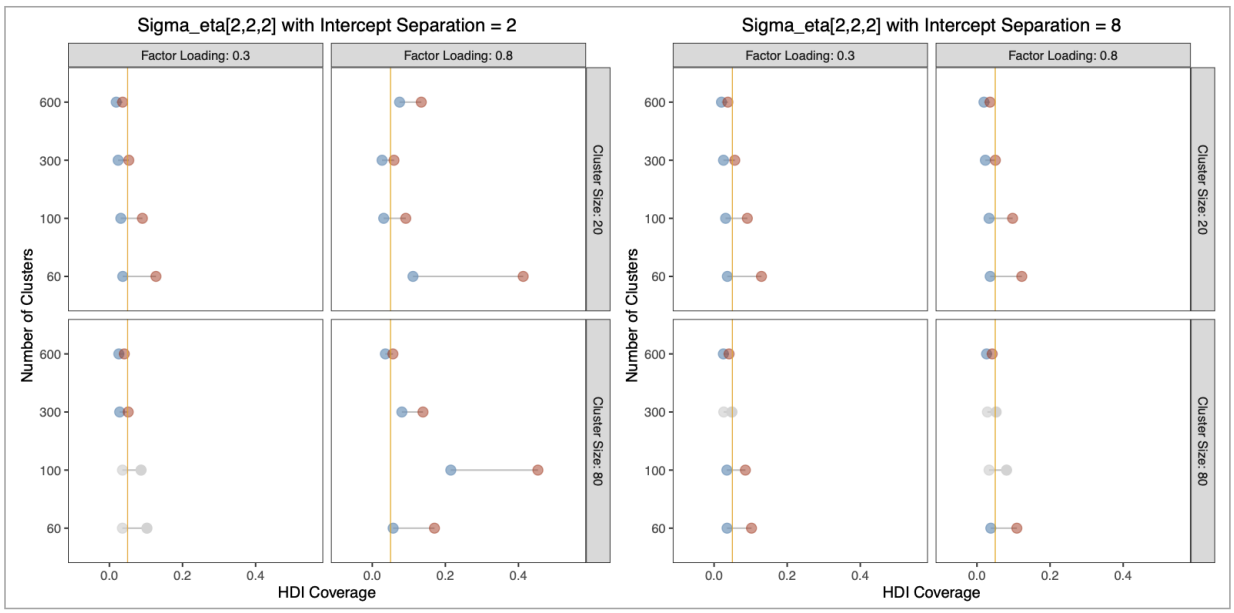
Note. (a) HPDI for estimating ϕ_{101}. (b) HPDI for estimating ϕ_{102}. Conditions that did not reach convergence are marked in light grey. Population value of the parameter is indicated by the golden line.

Figure 4-12

95% HPDI Coverage for Estimating Variance of the Within-Level Log Variance in Class 1 and Class 2



(a)



(b)

Note. (a) HPDI for estimating ϕ_{11}. (b) HPDI for estimating ϕ_{12}. Conditions that did not reach convergence are marked in light grey. Population value of the parameter is indicated by the golden line.

4.2.2.2 Naïve Mean Score MMSEM-RV

Based on the test results of Geweke's convergence diagnostic, the estimation of mean score based MMSEM-RV failed to reach convergence in only 3 out of 32 conditions (see Table B1 in Appendix B). Across the conditions that yielded converged solutions, the correct classification rate ranges from 0.89 to 1, with perfect classification results obtained in 15 conditions. There is high certainty in the classification decision—the values of CC_{mean} and CC_{min} are either equal to or approximately equal to 1 across all the converged conditions.

Although the naïve mean score based MMSEM-RV involves fewer model parameters compared to the latent construct based MMSEM-RV, they share the same focal parameters that are of inferential interest, given the specific purpose of modeling population heterogeneity in random variability. Again, the focal parameters would include class-specific between-level fixed effects $\boldsymbol{\gamma}_k^{(B)}$, class-specific between-level random effects covariance Φ_{vk}, and mixing proportions $\boldsymbol{\pi}$. The relative estimation bias for each of the focal parameters is presented in Figure B11 to Figure B15 in Appendix B. Across the conditions, only the mixing proportion ($\boldsymbol{\pi}$) is estimated with fairly good accuracy, with the magnitude of relative bias being within the range of .0006 to .265 (Figure B11). The other model parameters are estimated with much larger bias—particularly for the between-level variance and covariance parameters, for which the relative bias can go up to 8.775 (Figure B15). On average, the parameters are estimated with more severe bias for latent class 2, the sub-population that was generated with large within-level variability.

For most of the parameters, sample size does not appear to have an impact on estimation accuracy: increasing either the number of clusters or the cluster size does not help to reduce the estimation bias (with the only exception of ϕ_{1k} in class 1, for which the bias decreases as the number of clusters or cluster size goes up). Increased intercept separation, on the other hand,

seem to better contribute to the reduction in estimation bias. For all but one parameter (ϕ_{0k} in class 2), the estimation bias is noticeably smaller when the intercept is separated by 8 compared to the intercept separation of 2. With regard to measurement quality, higher factor loading (λ = 0.8) in the population model leads to lower average relative bias across all the model parameters except for ϕ_{1k} in class 2. Across the board, there is no single condition that shows satisfactory estimation accuracy for all the model parameters simultaneously.

The corresponding 95% HPDIs for estimating the focal parameters are plotted in Figures B16-B20 in Appendix B. Everything else being equal, the interval width gets narrower when the number of clusters *J* is 300 or 600, in comparison to the cases when *J* is at smaller values. The intercept separation and factor loading have no effect on the interval width for mixing proportion and between-level fixed effects. For the between-level random effect variance and covariance, a higher factor loading in the population model leads to slightly wider HPDI intervals, while larger item intercept separation yields narrower intervals. The cluster size does not have a noticeable impact on the interval width.

Regarding the interval coverage, the 95% HPDI can only capture the true value of mixing proportion π when the intercepts are separated by 8 in the population, or when the sample size is at its smallest when intercepts are separated by 2 (Figure B16). The 95% HPDI overall fails to cover the population value of the between-level fixed effects across all the conditions (Figure B17). The coverage is also not satisfactory for the between-level random effect variance (Figure B18)—while the interval is commonly off the target when factor loading is 0.3, the interval does slightly better at capturing the population value when the factor loading is 0.8 and the sample size is relatively small. With regard to the between-level covariance (Figure B19), the interval coverage looks better when intercept separation is 8 compared to when it is 2; however, the

interval in general fails to capture the true value when sample size is large. Similar observation is obtained for the variance parameter of the within-level random variability (Figure B20). Again, summarizing across all the conditions, there does not exist one single condition that yields satisfactory results regarding the interval coverage.

4.2.2.3 Comparison of Latent Construct MMSEM-RV and Naïve MMSEM-RV

To gain a better understanding of the performance differences between latent construct MMSEM-RV with a correctly specified measurement model as opposed to the naïve approach that is based on mean scores, relative bias for each of the focal parameters is compared side by side in Figure 4-13 to Figure 4-17.

In terms of the mixing proportion π, as shown in Figure 4-13, the naïve model yields substantially larger estimation bias when item intercepts are not widely separated ($\alpha = 2$); more specifically, the naïve model tends to underestimate the probability of belonging to latent class 1, the class that was generated with smaller within-level variability, and overestimate the probability of belonging to latent class 2, the class generated with a larger within-level variability. When item intercepts are well separated ($\alpha = 8$), however, the estimation bias for π is minimal under both approaches. Turning to the between-level fixed effects (Figure 4-14), the naïve mean score model considerably underestimates γ_{10k} for both class 1 and class 2—particularly when factor loading is low ($\lambda = 0.3$)—while latent construct MMSEM-RV yields much more accurate results across the conditions.

Regarding the between-level variance ϕ_{0k} (Figure 4-15), the naïve approach yields notably more negative bias than the latent construct MMSEM-RV when the intercept separation is at 8. When intercepts are separated by 2 and factor loading is 0.3, latent construct MMSEM-RV has

comparatively better performance in estimating ϕ_{0k} for latent class 1; however, when factor loading is 0.8 given the same intercept separation ($\alpha = 2$), neither latent construct MMSEM-RV nor naïve mean score MMSEM-RV demonstrates accurate point estimation. For the between-level random effect covariance, the naïve mean score approach consistently yields larger estimation bias than the latent construct model across all the conditions (Figure 4-16), particularly with small intercept separation. Lastly, with regard to the variance parameter of the within-level random variability (ϕ_{1k}), the naïve approach yields more estimation bias for ϕ_{1k} in class 2 when intercepts are separated by 2; the estimation accuracy of the two approaches are comparable for estimating ϕ_{1k} when intercept separation is 8 (Figure 4-17).

In summary, the naive mean-score model does not exhibit better or even comparable estimation accuracy across all focal parameters when contrasted with the latent construct MMSEM-RV. Overall, ignoring the measurement structure can lead to severe estimation bias. It is noted, however, that the naive approach can yield satisfactory point estimates for the mixing proportion when item intercepts are well separated. Additionally, acceptable results are obtained for estimating the random variability variance—and comparable to the results from latent construct model—when intercepts are well separated and sample size is large.

Figure 4-13

Relative Bias for Estimating the Probability of Belonging to Class 1 under the Two Different Modeling Approaches

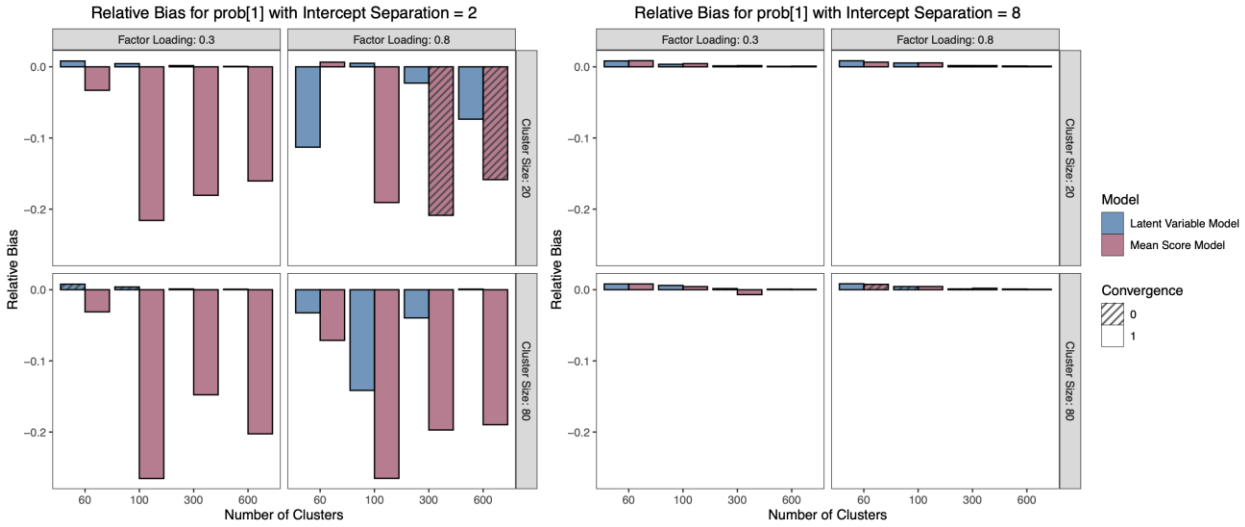
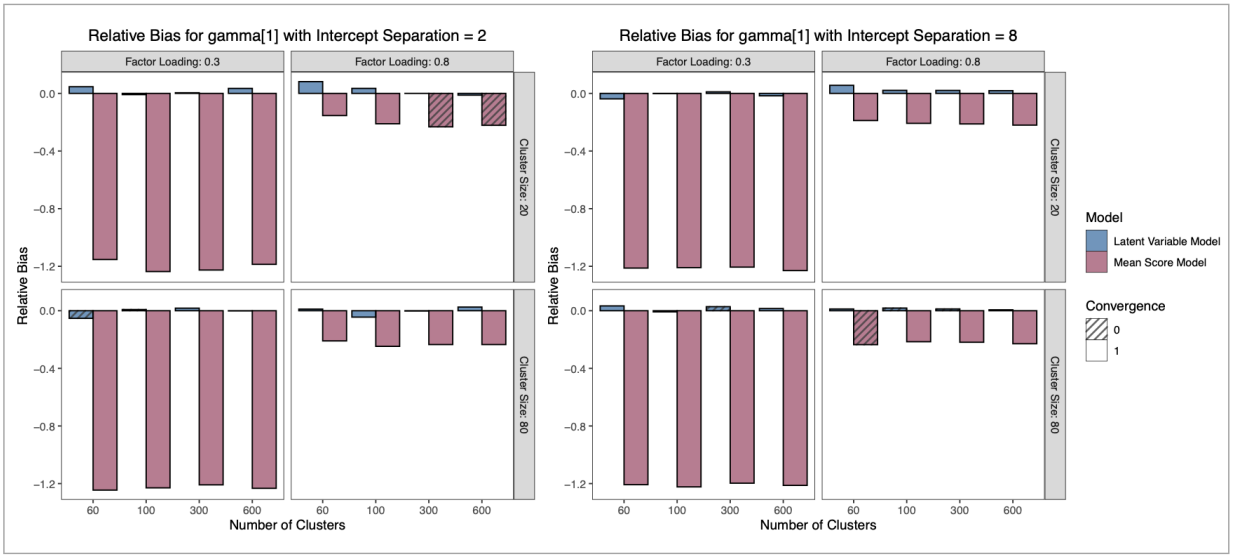
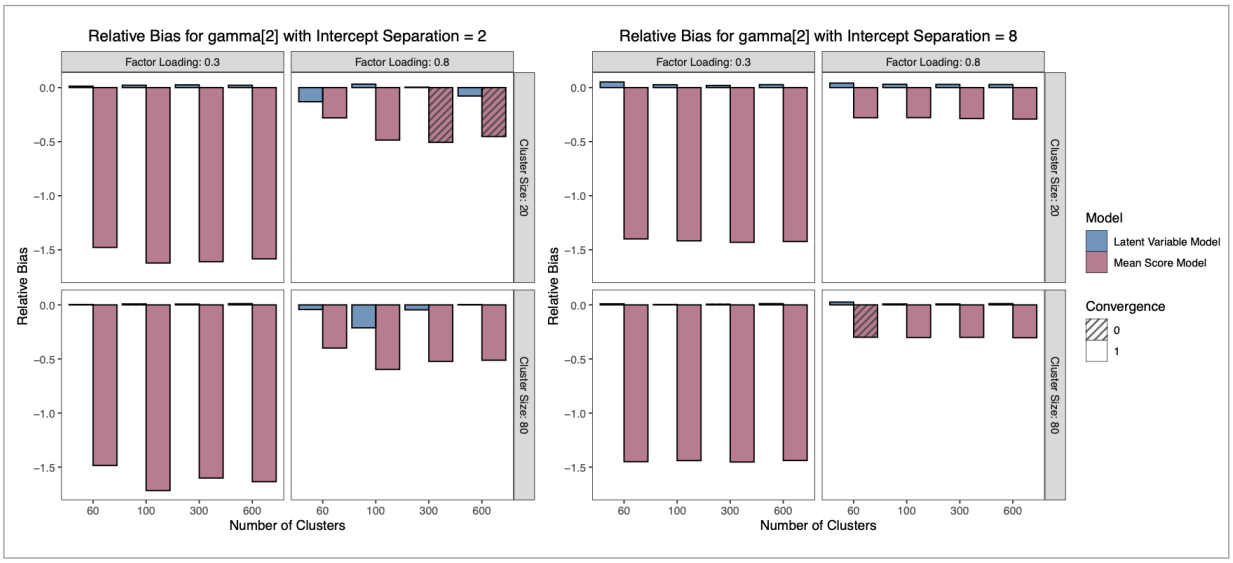


Figure 4-14

Relative Bias for Estimating Average Within-Level Log Variance in Class 1 and Class 2 under the Two Different Modeling Approaches



(a)

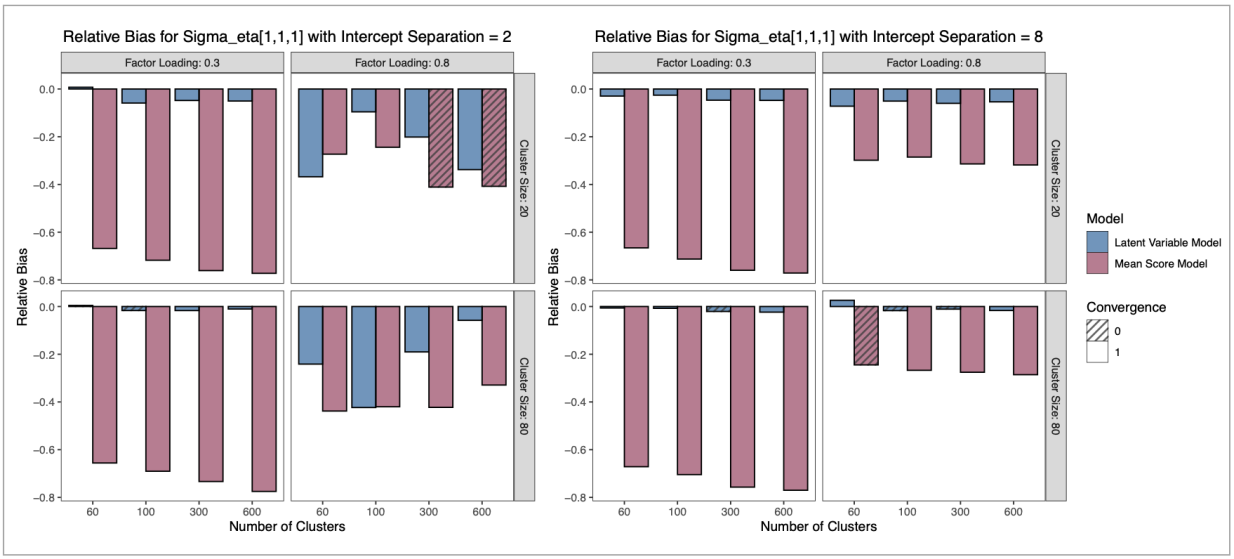


(b)

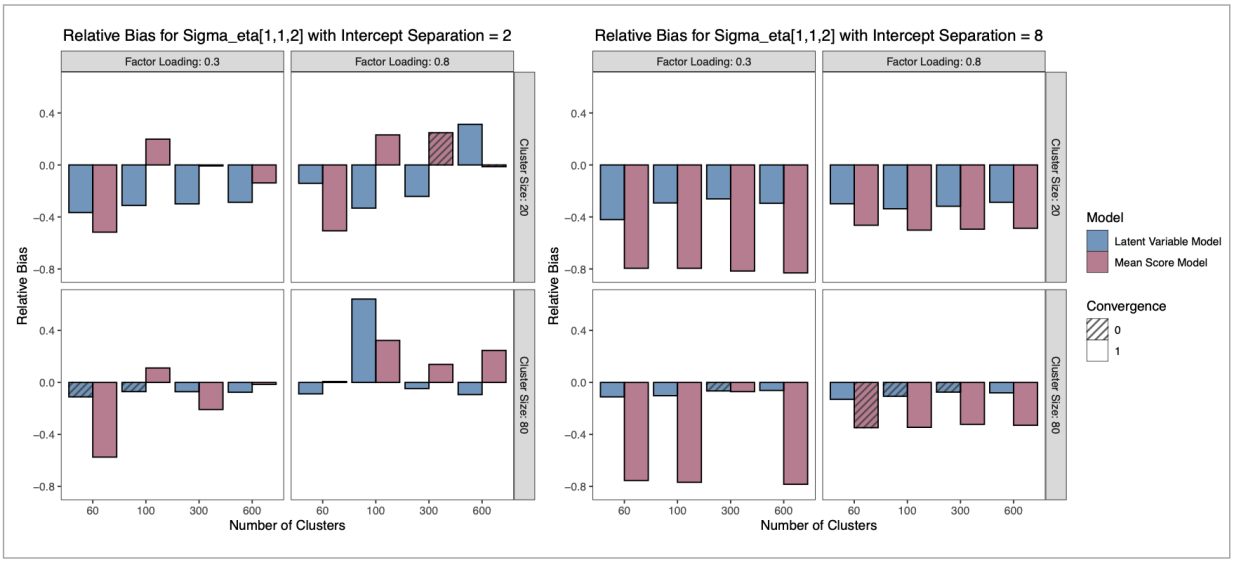
Note. (a) Relative bias for estimating γ_{101}. (b) Relative bias for estimating γ_{102}.

Figure 4-15

Relative Bias for Estimating the Between-Level Latent Construct Variance in Class 1 and Class 2 under the Two Different Modeling Approaches



(a)

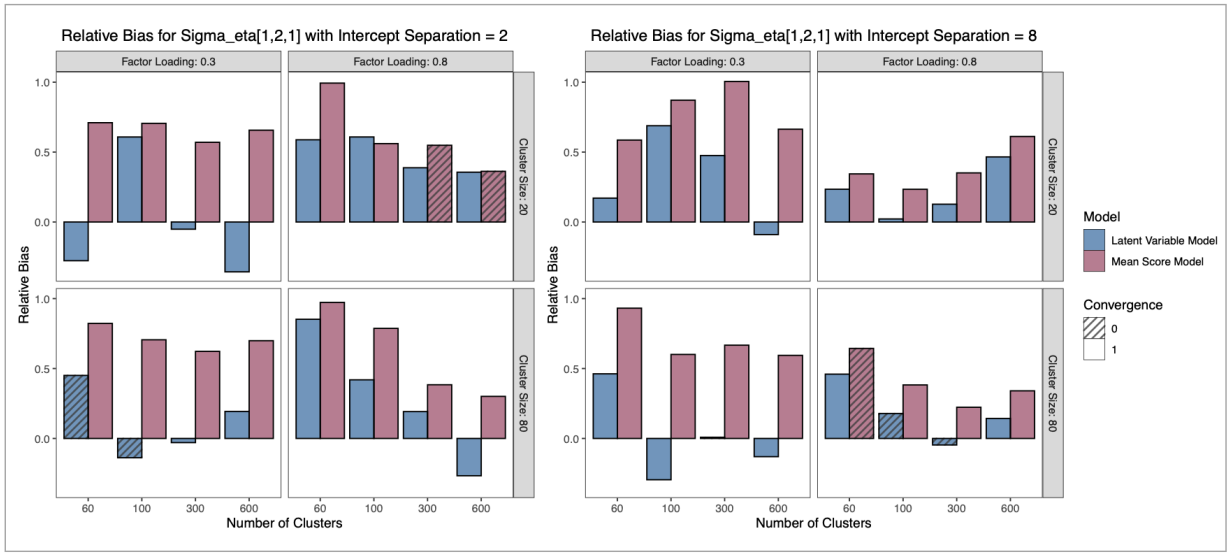


(b)

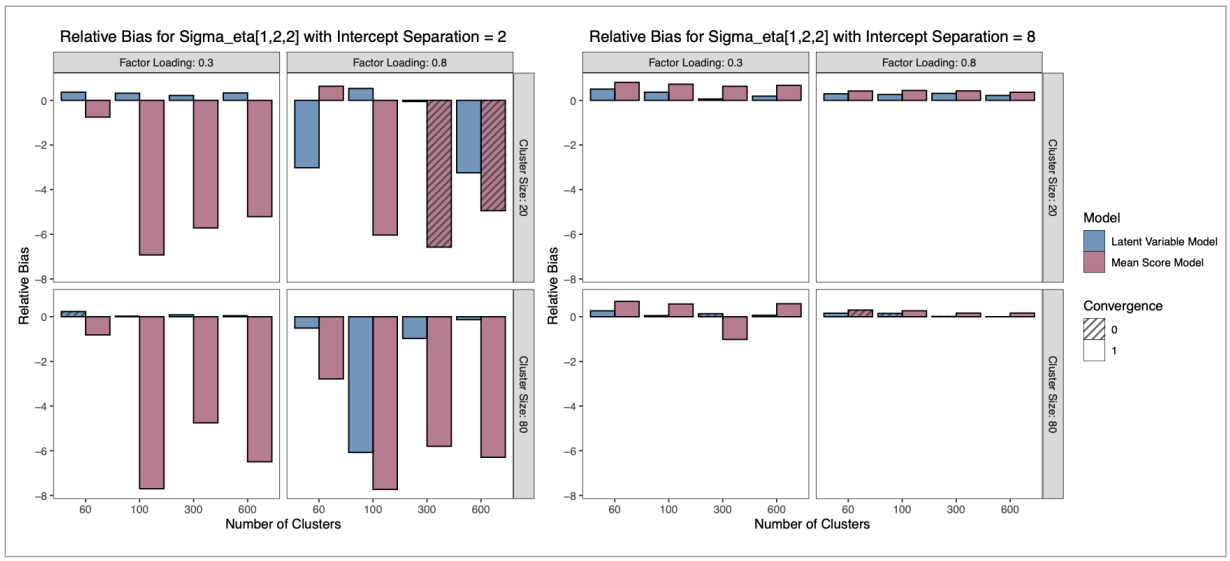
Note. (a) Relative bias for estimating ϕ_{01}. (b) Relative bias for estimating ϕ_{02}.

Figure 4-16

Relative Bias for Estimating the Between-Level Random Effects Covariance in Class 1 and Class 2 under the Two Different Modeling Approaches



(a)

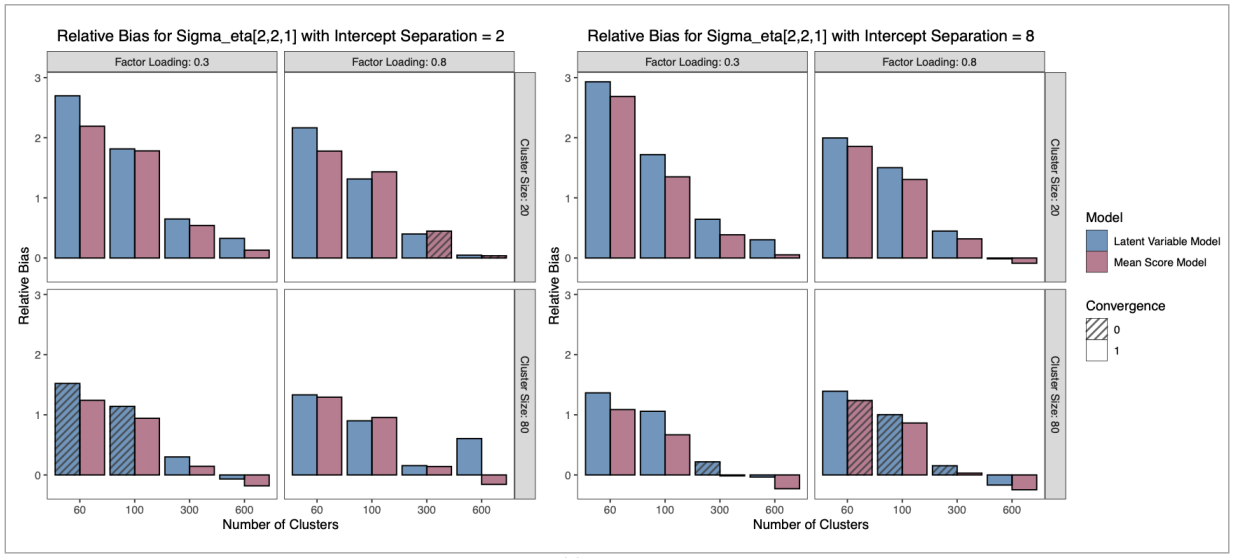


(b)

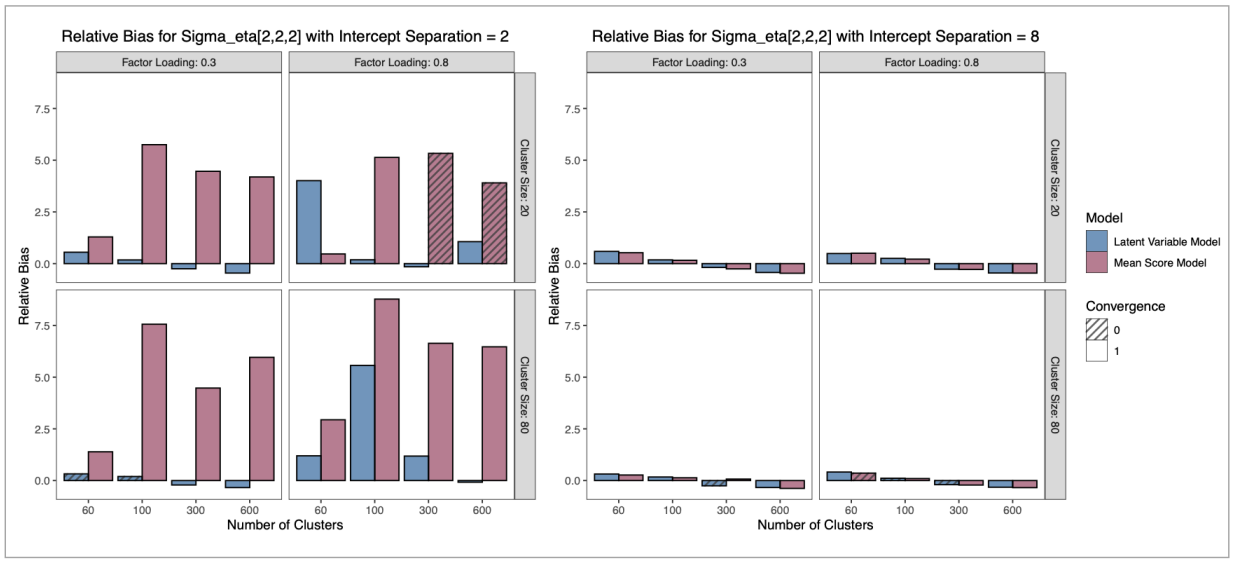
Note. (a) Relative bias for estimating ϕ_{101}. (b) Relative bias for estimating ϕ_{102}.

Figure 4-17

Relative Bias for Estimating Variance of the Within-Level Log Variance in Class 1 and Class 2 under the Two Different Modeling Approaches



(a)



(b)

Note. (a) Relative bias for estimating ϕ_{11}. (b) Relative bias for estimating ϕ_{12}.

4.3 Simulation Study Part II: Monte Carlo Simulation

In the second part of the stimulation study (part II), the performance of MMSEM-RV is evaluated using Monte Carlo methods under two selected conditions A ($J = 60$, $n_j = 20$, $\alpha_1 = 8$, $\lambda = 0.3$) and B ($J = 600$, $n_j = 80$, $\alpha_1 = 8$, $\lambda = 0.8$). In each of the conditions, 500 copies of simulated data were generated based on same the population data generating model specified previously. Similar as the population analysis, in each replication both the latent construct MMSEM-RV (Eq. (4.1)) and the naïve mean score based MMSEM-RV (Eq. (4.5)) with two latent classes ($k = 2$) were fit to the simulated data. The results were collected over the replications for further summary and discussion. Previous studies usually have the number of replications ranging from 100 to 500, which are deemed sufficient in the context of multilevel factor mixture models (e.g., Kim et al., 2016; Kim & Wang, 2019). The adequacy of the number of replications was further examined empirically in the current study (see results section below).

4.3.1 Outcome Evaluation

The simulation results were summarized over the replications within each of the two manipulated conditions respectively. For outcome evaluation, the following aspects were assessed and reported:

1) Convergence rate—the proportion of times that model estimation reached convergence.
2) Classification accuracy—the average correct class assignment rate, which is defined as the following:

$$\frac{\sum_{j=1}^{R}\sum^{J} I(C_j = \hat{C}_j)}{JR}, \tag{4.11}$$

where R is the total number of available replications within each condition.

3) Classification certainty—the average mean classification certainty (CC_{mean}) and average minimum classification certainty ($CC_{\min}$), which are defined as the following respectively:

$$\overline{CC}_{\text{mean}} = \frac{\sum_{j=1}^{R}\sum_{k=1}^{J}\sum_{k=1}^{K}\hat{C}_{jk}\tilde{P}(C_{jk})}{JR} \quad . \tag{4.12}$$

$$\overline{CC}_{\min} = \frac{\sum_{j}^{R}\min\sum_{k=1}^{K}\hat{C}_{jk}\tilde{P}(C_{jk})}{R}$$

4) Parameter estimation accuracy—the average relative bias for estimating each model parameter, comptuted as:

$$\frac{\left(\sum_{r=1}^{R}\hat{\theta}_r / R\right)-\theta}{|\theta|}. \tag{4.13}$$

5) Parameter estimation precision—both variance and root mean squared error (RMSE) of the parameter estimates are reported. The variance of parameter estimates across replications gauges to what extent the estimates vary from sample to sample, which is computed as:

$$s_\theta^2 = \frac{1}{R-1}\sum_{r=1}^{R}(\hat{\theta}_r - \bar{\theta})^2, \tag{4.14}$$

where $\bar{\theta}$ is the observed average point estimate across replications. RMSE, on the other hand, quantifies the estimation accuracy by taking into account both the bias and the variability of the estimates:

$$\text{RMSE} = \sqrt{\frac{\sum_{r=1}^{R}(\hat{\theta}-\theta)^2}{R}}. \tag{4.15}$$

An estimator with low RMSE thus indicate small bias and variance combined in parameter estimates.

6) HPDI coverage—the proportion of times when the 95% HPDI captures the population value of parameter θ:

$$\frac{\sum^{R} I\left[\text{HPDI}_{\text{LL}}(\hat{\theta}) \le \theta \le \text{HPDI}_{\text{UL}}(\hat{\theta})\right]}{R}. \tag{4.16}$$

4.3.2 Results

The Monte Carlo simulation study was implemented using R via the University of Maryland's HPC cluster Zaratan (in total 1000 cores were requested for the Monte Carlo simulation jobs, while a maximum of 500 cores were requested to be used simultaneously at one time). It took about 17 minutes to fit one latent construct MMSEM-RV model in condition A (J = 60, n_j = 20, α_1 = 8, λ = 0.3) and about 58 hours to fit the same model in condition B (J = 600, n_j = 80, α_1 = 8, λ = 0.8). Comparatively, the running time for fitting a naïve mean score MMSEM-RV is about 1.9 minutes in condition A and about 2.9 hours in condition B. As similarly found in the population analysis, the computational time increases tremendously as the sample size increases. Also as expected, the naïve mean score model is less computational demanding.

4.3.2.1 Simulation Condition A

According to the test results of Geweke's convergence diagnostic, the latent construct MMSEM-RV reached convergence in 477 out of the 500 replications (convergence rate = .954) while the naïve mean score model reached convergence in 496 replications (convergence rate = .992) under condition A (J = 60, n_j = 20, α_1 = 8, λ = 0.3). Only the results obtained from converged replications are included for outcome evaluation and the non-converged replications

are discarded. To assess whether the remaining number of replications are sufficient, the running mean plots for the parameter point estimates over the number of converged replications is examined (Figure B21 and Figure B22 in Appendix B). It shows that the point estimates for the model parameters stabilized over the available replications in the current study.

Across the converged replications, the latent construct MMSEM-RV yields an average classification accuracy of 1, indicating perfect classification in each replication; the naïve mean score model also yields an average correct classification rate that approximates 1, suggesting perfect classification in most replications. For both approaches, the classification decision is made with high certainty—with the value of $\overline{CC}_{\text{mean}}$ and $\overline{CC}_{\text{min}}$ both being equal to 1.

The estimation results are closely examined for the focal model parameters $\boldsymbol{\gamma}_k^{(B)}$, Φ_{vk}, and $\boldsymbol{\pi}$. Plots of the average point estimates with Monte Carlo standard error are provided in Figure 4-18, with Monte Carlo standard error s_{MC} defined as

$$s_{\text{MC}} = \frac{\sqrt{\sum(\hat{\theta} - \bar{\theta})^2}}{\sqrt{R_c - 1}\sqrt{R_c}}, \tag{4.17}$$

where R_c denotes the number of converged replications. As shown in the figures, with the latent construct MMSEM-RV, the point estimate is only accurate for the fixed effect γ_{102} and between-level latent construct variance ϕ_{01}; the estimate is slightly downward biased for γ_{101} and slightly upward biased for π_1, while greater bias is observed for other between-level random effect variance covariance parameters, including ϕ_{101}, ϕ_{11}, ϕ_{02}, ϕ_{102}, and ϕ_{12}. The model tends to overestimate the between-level covariance and the variance of the random within-level variability in both classes; but it underestimates the between-level construct variance in class 2 (also see relative bias provided in Figure 4-19). In contrast to the latent construct MMSEM-RV, the naïve mean score model yields biased results for all the focal parameters. Particularly, for the

fixed effects (γ_{101}, γ_{102}) and between-level variance and covariance (ϕ_{01}, ϕ_{101}, ϕ_{02}, and ϕ_{102}), the naïve approach yields substantially larger bias than the latent construct model. Regardless of the modeling approaches, the estimation of the variance of random variability appears to suffer from most severe bias under this condition.

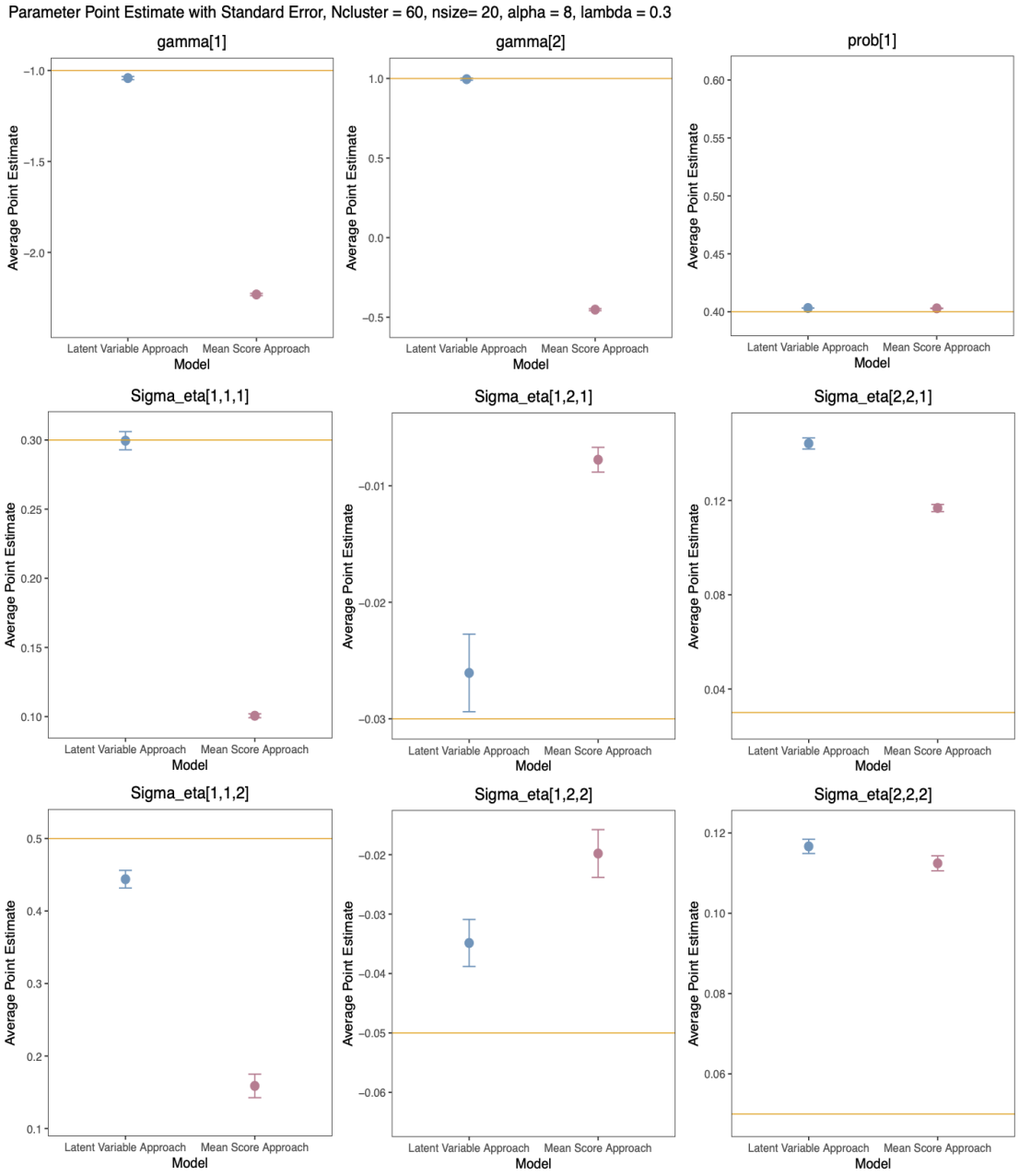
The 95% HPDI coverage probability for the focal model parameters is shown in Figure 4-20. With latent construct MMSEM-RV, the coverage of 95% HPDI reaches the nominal level (0.95) for most of the focal parameters, although the coverage is less than satisfactory for ϕ_{101} ϕ_{102}, and ϕ_{12}. On the other hand, with the naïve approach, the 95% HPDI fails to achieve the expected coverage probability for all but two model parameters. For several parameters (e.g., the fixed effects), the coverage probability of the 95% HPDI obtained with the naïve approach is zero, suggesting extremely poor performance. Across the focal parameters, the coverage probability is least satisfactory for the variance of random variability under this condition.

In terms of the estimation precision, the Monte Carlo variance across replications is reported for the focal parameters (Table 4-3). Both modeling approaches demonstrate minimal variation in the point estimates from sample to sample. The RMSE for focal parameters is also reported in Table 4-3. Comparing latent construct MMSEM-RV to the mean score model, the latter results in substantially larger RMSE in estimating the fixed effects (γ_{101}, γ_{102}) and the between-level latent construct variance (ϕ_{01}, ϕ_{02}).

In summary, under the condition with small sample size, low factor loading, and large intercept separation ($J = 60$, $n_j = 20$, $\alpha_1 = 8$, $\lambda = 0.3$), the latent construct MMSEM-RV generally outperforms the naïve mean score approach. However, the latent model fails to yield satisfactory results for estimating the variance of variability—which is deemed most challenging in prior population analysis.

Figure 4-18

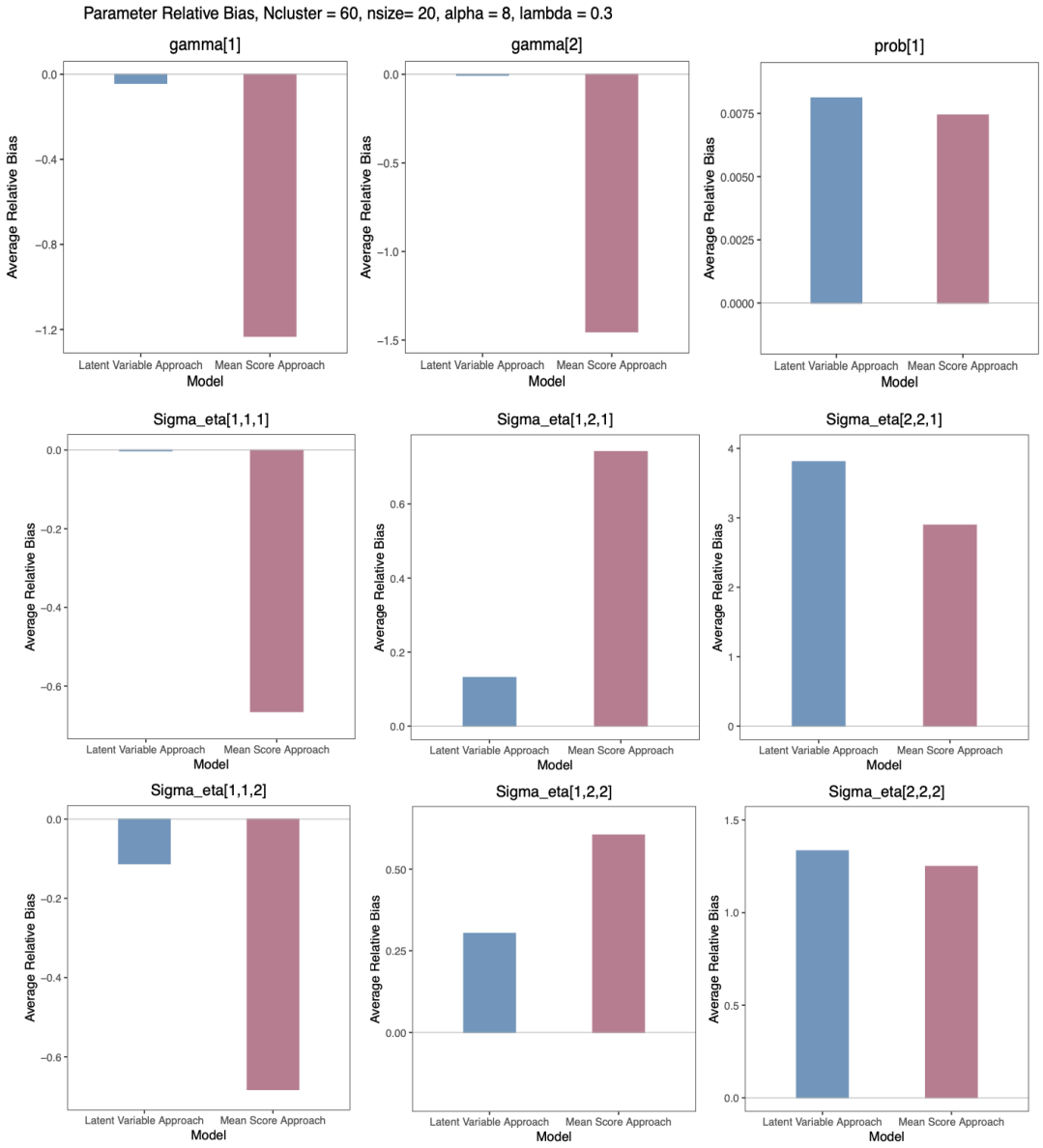
Average Point Estimates with Standard Error under Condition A



Note. The population true values are indicated by the golden line. The error bars are computed as $\bar{\theta} \pm 1.96 \times s_{MC}$. The model parameters are arranged in the following order for this graph from left to right by each row: $\gamma_{101}, \gamma_{102}$, π_1, ϕ_{01}, ϕ_{101}, ϕ_{11}, ϕ_{02}, ϕ_{102}, ϕ_{12}.

Figure 4-19

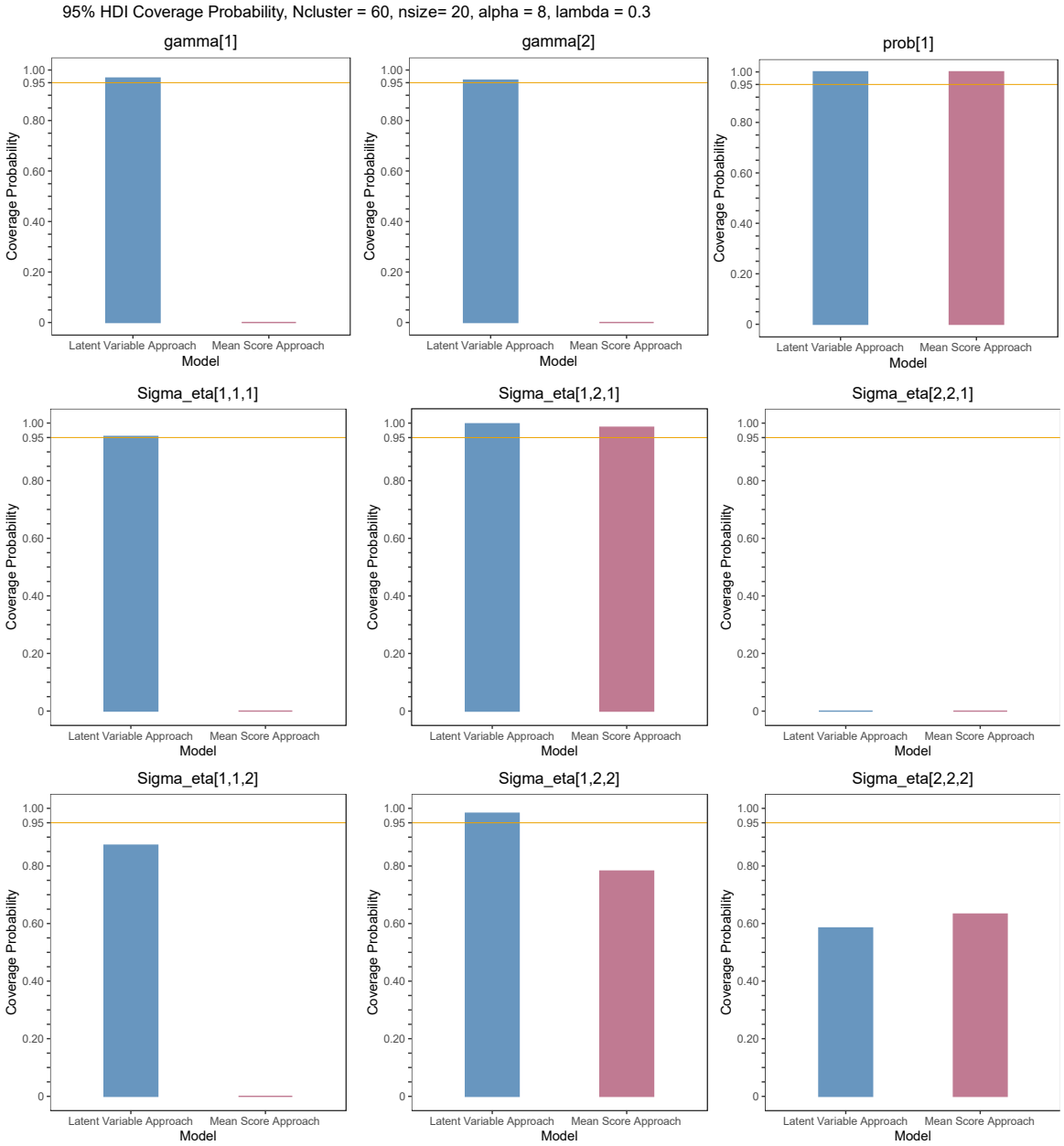
Relative Bias for Focal Parameters under Condition A



Note. The model parameters are arranged in the following order for this graph from left to right by each row: γ_{101}, γ_{102}, π_1, ϕ_{01}, ϕ_{101}, ϕ_{11}, ϕ_{02}, ϕ_{102}, ϕ_{12}.

Figure 4-20

95% HPDI Coverage Probability for Focal Parameters under Condition A



Note. The model parameters are arranged in the following order for this graph from left to right by each row: γ_{101}, γ_{102}, π_1, ϕ_{01}, ϕ_{101}, ϕ_{11}, ϕ_{02}, ϕ_{102}, ϕ_{12}. The target coverage probability (95%) is indicated by the golden line.

Table 4-3

Monte Carlo Variance and RMSE Obtained from the Monte Carlo Simulation

Focal Parameter	Condition A				Condition B			
	Latent MMSEM-RV		Naïve MMSEM-RV		Latent MMSEM-RV		Naïve MMSEM-RV	
	Monte Carlo Variance	RMSE	Monte Carlo Variance	RMSE	Monte Carlo Variance	RMSE	Monte Carlo Variance	RMSE
γ_{101}	.009	.106	.005	1.233	2.54E-04	.016	2.07E-04	.235
γ_{102}	.006	.074	.005	1.454	2.06E-04	.014	2.29E-04	.315
π_1	7.19E-08	.003	4.77E-06	.004	8.85E-09	3.47E-04	1.32E-06	.001
ϕ_{01}	.005	.073	2.62E-04	.200	.001	.027	4.01E-04	.085
ϕ_{101}	.001	.037	1.45E-04	.025	.000	.009	4.81E-05	.010
ϕ_{11}	.001	.117	3.00E-04	.089	1.84E-05	.012	1.23E-05	.007
ϕ_{02}	.018	.147	.034	.388	.002	.040	.011	.123
ϕ_{102}	.002	.047	2.09E-03	.055	1.29E-04	.011	.001	.027
ϕ_{12}	3.92E-04	.070	4.50E-04	.066	2.92E-05	.007	5.89E-05	.011

4.3.2.2 Simulation Condition B

Based on the results of Geweke's convergence diagnostic, the convergence rate of latent construct MMSEM-RV is .640 out of 500 replications (R_c = 320), while the naïve mean score model yields a convergence rate of .976 (R_c = 488) under condition B (J = 600, n_j = 80, α_1 = 8, λ = 0.8). The replications that failed to reach convergence are not included for further evaluation. Running mean plots for parameter estimates over the number of available replications were visually inspected to assess the sufficiency of replications (Figure B23 and Figure B24 in Appendix B). As shown in the plots, the point estimates for all model parameters stabilized fairly well over the converged replications under the current simulation condition.

With regard to classification accuracy, the latent construct MMSEM-RV yields an average classification accuracy of 1, obtaining perfect cluster classification in all 320 replications; the naïve mean score model yields an average correct classification rate that approximates 1 across 488 replications, with perfect cluster classification obtained in 289 replications. For both approaches, the classification decision is made with high certainty—the $\overline{CC}_{\text{mean}}$ and $\overline{CC}_{\text{min}}$ are both equal to 1.

To assess the performance of parameter estimation, let us continue to focus on the focal model parameters $\boldsymbol{\gamma}_k^{(B)}$, $\boldsymbol{\Phi}_{vk}$, and $\boldsymbol{\pi}$. The average point estimates with Monte Carlo standard error are displayed in Figure 4-21. As shown in the plots, latent construct MMSEM-RV yields accurate point estimates for the fixed effects $\boldsymbol{\gamma}_k^{(B)}$ and negligible bias for estimating mixing proportion π_1 (with a relative bias of .0008). As for the between-level variance and covariance, the point estimates are unbiased for the latent construct variances ϕ_{0k} and covariances ϕ_{10k}. The point estimates are, however, slightly upward biased for the variance of the within-level

variability, with relative bias of .377 for ϕ_{11} in class 1 and .103 for ϕ_{12} in class 2 (see Figure 4-22). On the other hand, as opposed to latent construct MMSEM-RV, the naïve mean score model yields biased estimates for all the focal parameters. More specifically, it underestimates the fixed effects $\gamma_k^{(B)}$, mixing proportion π_1, and latent construct variances ϕ_{0k} (see Figure 4-22), while overestimating the variance of variability ϕ_{1k}. Regarding the covariance, the estimate is positively biased for ϕ_{101} in class 1, but negatively biased for ϕ_{102} in class 2.

The 95% HPDI coverage probability for the focal model parameters is provided in Figure 4-23. As shown in the plots, the 95% HPDI has satisfactory coverage performance for most of the focal parameters when latent construct MMSEM-RV is fit to the data, with the only exception being the variance of within-level variability ϕ_{1k}. In contrast, with the naïve mean score approach, the 95% HPDI fails to reach the expected 95% coverage probability for all focal parameters except the mixing proportion π_1. The coverage performance is particularly poor for the fixed effects $\gamma_k^{(B)}$ with the naïve approach; the coverage probability is essentially zero, indicating that the interval estimation is overall off target.

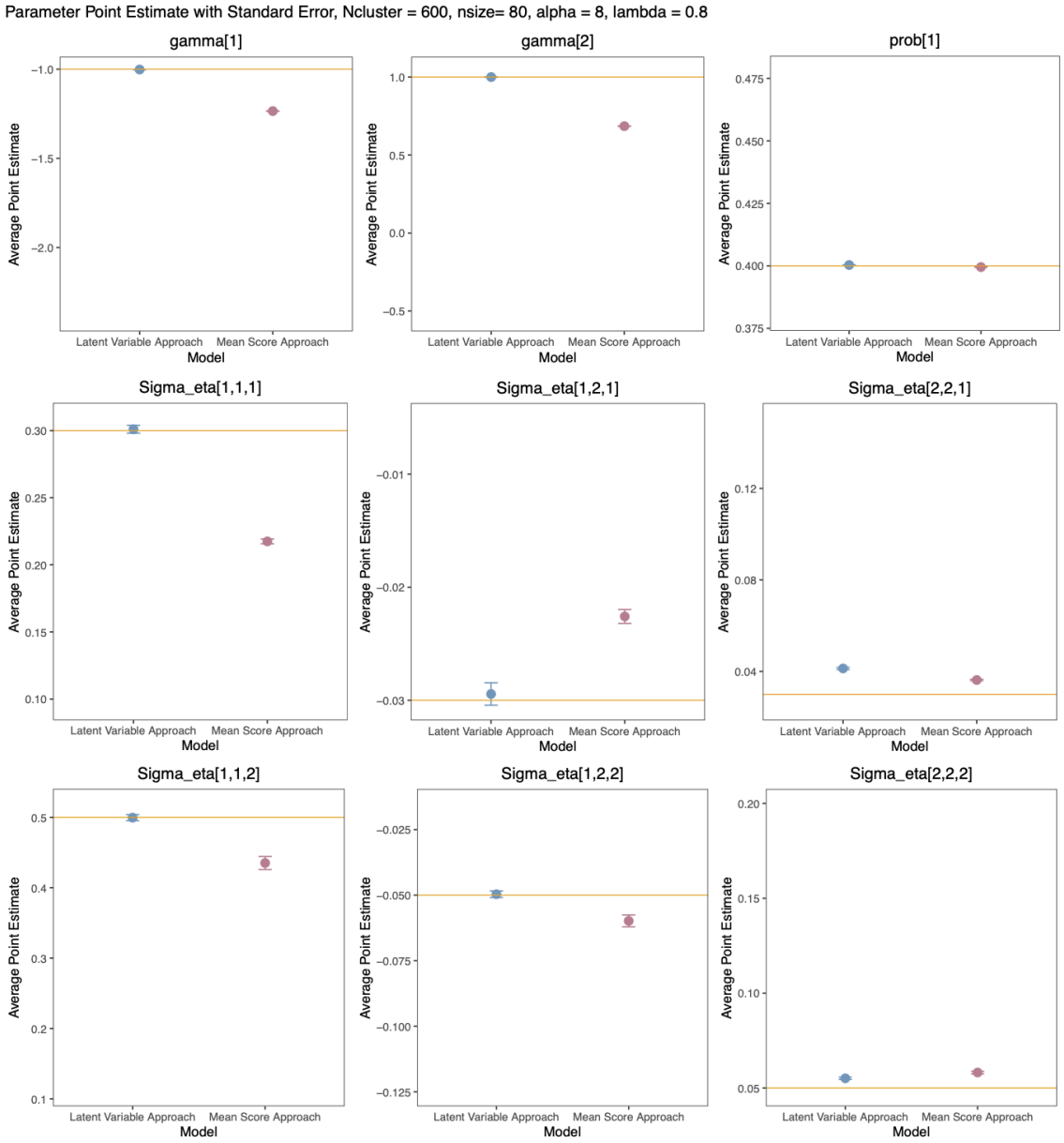
In terms of the estimation precision, the Monte Carlo variance across replications and RMSE are both reported for the focal parameters (Table 4-3). Similarly as in condition A, the point estimates have minimal variation from sample to sample under both modeling approaches. With regard to RMSE, the latent construct MMSEM-RV yields substantially smaller RMSE for estimating most of the focal parameters than the naïve mean score model does.

To summarize, under the condition of large sample size, high factor loading, and large intercept separation ($J = 600$, $n_j = 80$, $\alpha_1 = 8$, $\lambda = 0.8$), the latent construct MMSEM-RV outperforms the naïve mean score approach in terms of estimation accuracy, precision, as well as HPDI coverage probability. However, the estimation of variance of variability again appears to

be most challenging under this condition. Neither approach yields satisfactory results for this model parameter under the current condition, although they are much better compared to the ones obtained under condition A.

Figure 4-21

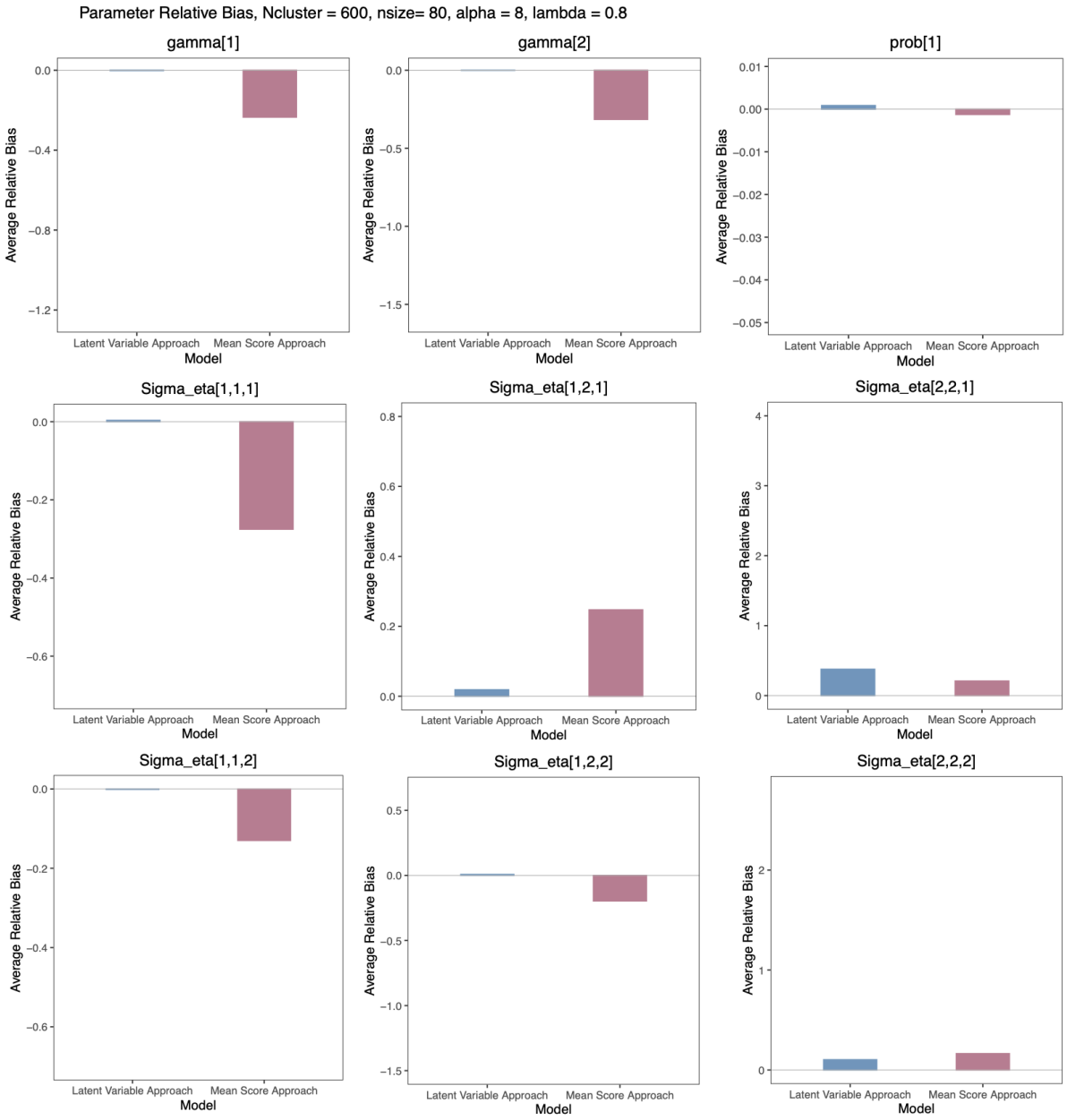
Average Point Estimates with Standard Error under Condition B



Note. The population true values are indicated by the golden line. The error bars are computed as $\bar{\theta} \pm 1.96 \times s_{MC}$. The model parameters are arranged in the following order for this graph from left to right by each row: $\gamma_{101}, \gamma_{102}$, π_1, ϕ_{01}, ϕ_{101}, ϕ_{11}, ϕ_{02}, ϕ_{102}, ϕ_{12}.

Figure 4-22

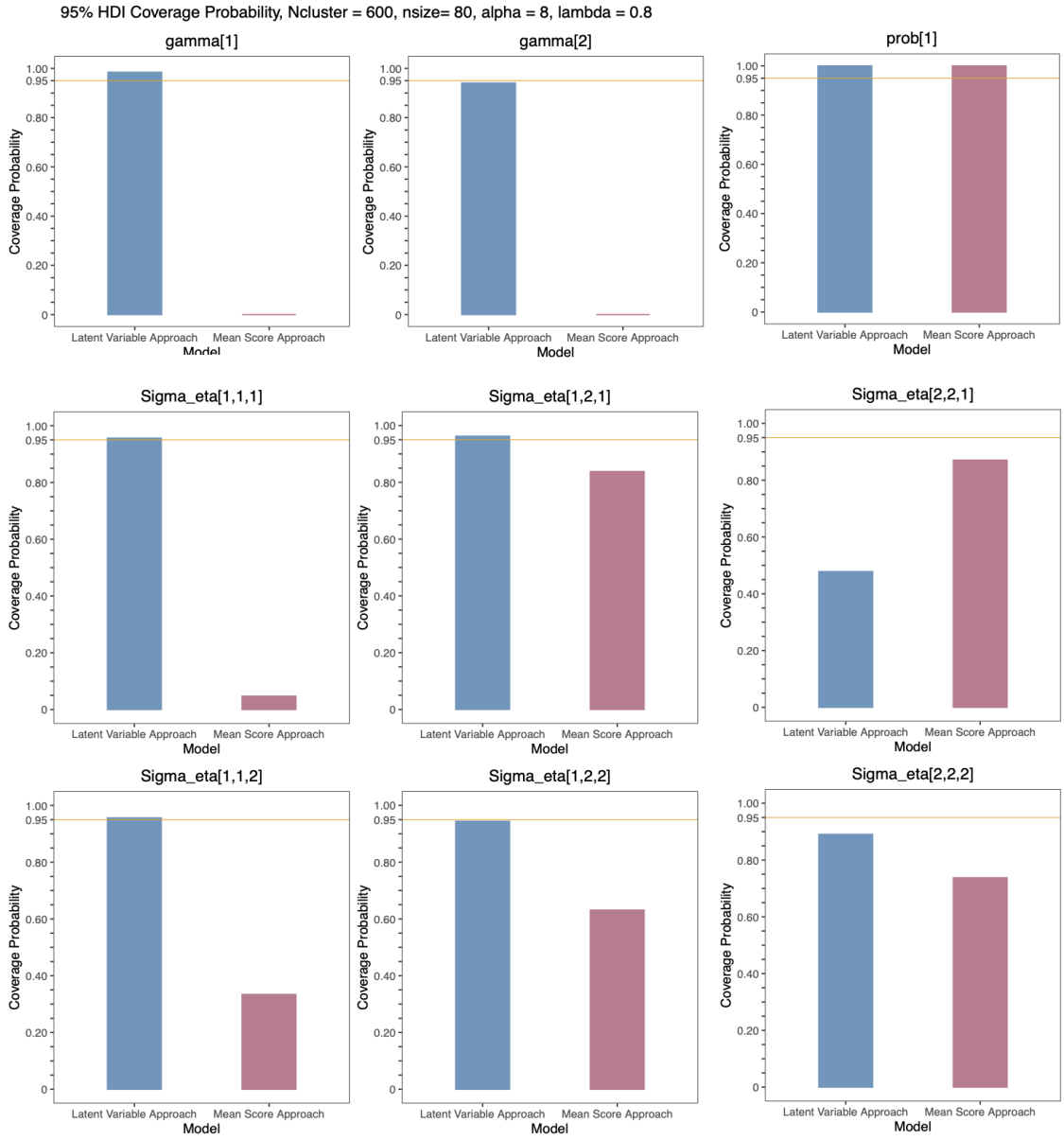
Relative Bias for Focal Parameters under Condition B



Note. The model parameters are arranged in the following order for this graph from left to right by each row: $\gamma_{101}, \gamma_{102}$, π_1, ϕ_{01}, ϕ_{101}, ϕ_{11}, ϕ_{02}, ϕ_{102}, ϕ_{12}.

Figure 4-23

95% HPDI Coverage Probability for Focal Parameters under Condition B



Note. The model parameters are arranged in the following order for this graph from left to right by each row: $\gamma_{101,}\gamma_{102}$, π_1, ϕ_{01}, ϕ_{101}, ϕ_{11}, ϕ_{02}, ϕ_{102}, ϕ_{12}. The target coverage probability (95%) is indicated by the golden line.

4.4 Summary of Simulation Study Results

Through a combination of population analysis and Monte Carlo simulation, the performance of MMSEM-RV is systematically evaluated across various conditions. The primary findings and main conclusions for the core research questions (RQ) are summarized below.

RQ1. Can MMSEM-RV be effectively estimated using a Bayesian approach via Gibbs sampling, yielding a converged solution? Bayesian estimation via Gibbs sampling is shown to be a viable and effective approach for estimating MMSEM-RV. Although no existing software yet offers the built-in Bayesian estimation for MMSEM-RV, it is straightforward to specify the model syntax in BUGS language (see Appendix C) and have the model estimated in R. The computational burden increases as the sample size goes up, and the model convergence also gets more challenging with larger sample size. In such cases, longer chains with more iterations may be needed for better model convergence.

RQ2. Can the focal model parameters of MMSEM-RV be accurately and efficiently estimated? The estimation accuracy varies across different model parameters given the same simulated condition. Some model parameters are estimated more accurately than others. For instance, among all the focal parameters, minimal estimation bias is observed for between-level fixed effects, as well as for the mixing proportions. The bias appears to be more salient for between-level variances and covariances across most of the conditions—particularly for the variance parameter of within-level variability, which is in general overestimated. To achieve better estimation accuracy for these parameters, a substantially larger sample size may be required. With regard to estimation precision, the estimates obtained under the simulation study are pretty consistent across repeated sampling, showing fairly small variation.

RQ3. How accurate is the latent class assignment when MMSEM-RV is used for classifying clusters, and how uncertain is the classification decision being made? The proposed MMSEM-RV demonstrates remarkably satisfactory classification accuracy across the various conditions in the simulation study. Employing Bayesian estimation, the classification decisions are not only accurate, but also made with a high degree of certainty.

RQ4. To what extent do factors such as sample size, class separation, and measurement quality impact the estimation of MMSEM-RV? A large sample size (particularly a large number of clusters) can effectively reduce the estimation bias. Given the same level of class separation in latent variability, a wider separation in item intercepts can also considerably reduce estimation bias, especially when the factor loadings are high. The estimation accuracy is more problematic with small intercept separation between latent classes while the observed measures are of good composite reliability. Interestingly, the classification accuracy and classification certainty are not sensitive to the varying factors. Among the focal model parameters, the estimation of between-level fixed effects and mixing probability is also more robust in face of less favorable conditions compared to other parameters.

RQ5. How does the proposed MMSEM-RV model compare to a naïve mean score based approach that ignore the measurement error? Compared to latent construct MMSEM-RV with a correctly specified measurement model, the naïve mean score approach in general yields larger bias, worse interval coverage, and less precision in parameter estimates. Across all simulated conditions, the estimation outcomes of the naïve approach consistently fail to render satisfactory results. Notably, in no single condition does the naïve approach yield satisfactory estimation results for all focal parameters simultaneously. With that being said, however, the naïve

approach does have good classification accuracy. It also shows low bias for estimating the mixing probability when item intercepts are widely separated between latent classes.

Chapter 5. Applied Illustrative Example

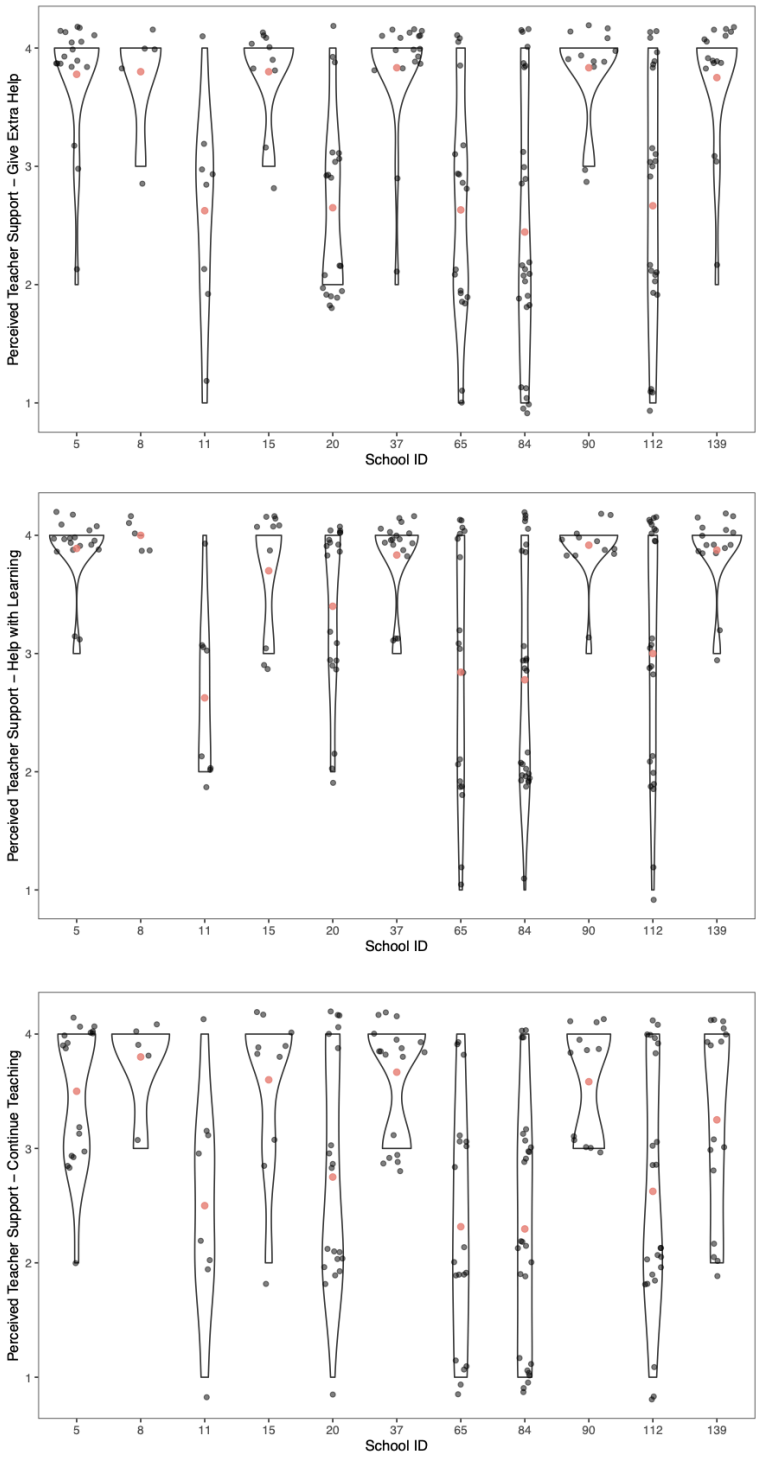
5.1 Research Context and Analytical Data

To demonstrate the implementation of MMSEM-RV in practice, an illustrative example involving empirical data is presented in this chapter. The illustrative example employs an unconditional MMSEM-RV to detect potential underlying sub-populations that have different distributions of within-school variability in student perceptions. The example utilizes the publicly available data from the Programme for International Student Assessment (PISA), a large-scale international study administered by the Organisation for Economic Co-operation and Development (OECD). More specifically, the U.S. data from PISA 2012 (OECD, 2013a, 2013b, 2013c) is used for this illustrative example, where the outcome of focal interest is the student perceived support from math teachers. The analytic dataset is downloaded via the `EdSurvey` R package (Bailey et al., 2023), which contains the cross-sectional data from 3,232 students (1,580 females; Mean age = 15.52 years, SD = 0.5) in 162 schools. The average cluster size is $\bar{n}_j$ = 19.95 across the participating schools. Three items each asking how often a described scenario occurs in math classes are used as observed indicators of latent perceived support: 1) "The teacher gives extra help when students need it," 2) "The teacher helps students with their learning," and 3) "The teacher continues teaching until the students understand.". Students responded to the items on a 4-point perceived frequency scale (1 = every lesson; 4 = never or hardly ever). All items are reverse coded prior to data analyses such that higher values correspond to higher levels of perceived support.

The distribution of the three observed indicators in a random sample of 11 schools is presented in Figure 5-1. In addition to the discrepancies in the mean levels of student perceived

Figure 5-1

Distribution of the Three Items Measuring Perceived Math Teacher Support within Each School for a Random Sample of 11 Schools



Note. The cluster averages are indicated by the red dot in each school.

teacher support, there are also notable differences among schools in terms of the within-cluster variability. As shown in the plots, while students in some schools share similar perceived levels of support, students in other schools demonstrate more heterogeneous perceptions regarding the support received from math teachers. For instance, school 8 and school 90 appear to be a lot more homogeneous than school 84 and 112 in terms of students' perspectives. The observed data pattern suggests the possibility that there can be unobserved grouping of schools in the broader population. However, it is unclear to what extent the observed heterogeneity in manifest indicators is inflated by measurement error. In order to detect the potential sub-populations of schools that have varying levels of within-school heterogeneity while taking measurement error into consideration, a latent construct MMSEM-RV is fit to the analytical data.

5.2 Methods

To simplify the demonstration, the item responses are treated as continuous variables in the current example[8]. Assuming the assumptions discussed in Chapter 3 are satisfied, two unconditional latent construct MMSEM-RV models are fit to the analytical data set, one with two latent classes ($K = 2$) and the other with three latent classes ($K = 3$). For both models, the parameteres are estimated using Bayesian approach, implemented via Gibbs sampling using the R package `rjags` (Plummer et al., 2022). Both models are estimated with one single chain and

[8] Ideally, the 4-point perceived frequency responses should be treated as ordinal variables. It is straightforward to generalize MMSEM-RV to accommodate ordinal variables by assuming an underlying latent continuum. However, it will inevitably increase the computational burden and may result in potential convergence issues given the complexity of the model and the current sample size. Therefore, for simplicity of illustration purposes, as well as to reduce the number of parameters to be estimated, this example treated the 4-point item responses as continuous variables. The results of this current example thus should not be interpreted beyond its illustrative purpose of applying MMSEM-RV with empirical data.

a total number of 50,000 iterations after burn-in. The priors are specified similarly as those employed in the simulation study (

Table ***5-1***), with necessary adjustments based on the observed sample statistics of the analytical PISA data. The order constraints $\gamma_{101} < ... < \gamma_{10K}$ are imposed on the fixed effects to avoid label switching during the iterations.

Table 5-1

Parameter Priors Utilized in the Applied Illustrative Example

Parameter	2-Class Model	3-Class Model
γ_{101}	N(-0.3, 100000)	N(-0.4, 100000)
γ_{102}	N(0, 100000)	N(-0.15, 100000)
γ_{103}	-	N(0, 100000)
$\boldsymbol{\pi}$	Dir($\boldsymbol{\alpha}$ = (1, 1))	Dir($\boldsymbol{\alpha}$ = (1, 1, 1))
$\boldsymbol{\Phi}_{vk}$	W^{-1}($\mathbf{I}_2$, 6)	W^{-1}($\mathbf{I}_2$, 6)
α_1	N(3.4, 10)	N(3.5, 10)
α_2	N(2.6, 10)	N(3.2, 10)
α_3	-	N(2.6, 10)
λ_i	N(0, 10000000)	N(0, 10000000)
$\boldsymbol{\Theta}_{iik}^{(W)}$	Inv-Gamma(3, 0.5)	Inv-Gamma(3, 0.5)

Geweke's diagnostic (Geweke, 1992) is utilized to assess model convergence, following the same procedure outlined in Chapter 4. Additionally, the trace plots and running mean plots for each model parameter are also visually inspected for further convergence diagnostic. The model is considered converged only if all the model parameters reaches convergence after 50,000 iterations. Once passing the convergence test, the posterior distributions of model parameters are examined, with posterior means reported as the point estimates and the 95%

HPDI reported as the interval estimates. For school classification, the posterior mode of the class assignment for each school *j* is recorded as the predicted class membership $\hat{C}_j$.

Once converged, the 2-class and 3-class models are compared against each other using a collection of information criteria, including AIC (Akaike information criterion), BIC (Bayesian information criterion), sample size adjusted BIC (saBIC), and IC1000. For each information criterion, the model that yields a lower value is preferred as a better fit. These information criterion are formally defined as a function of the model deviance $D(\mathbf{y},\boldsymbol{\theta})$ in the literature. In the current study, to assess model fit, the posterior distribution of deviance is monitored over the iterations, and the posterior mean deviance is found after the model converges:

$$\bar{D}(\mathbf{y},\boldsymbol{\theta}) = E\left[D(\mathbf{y},\boldsymbol{\theta})\middle|\mathbf{y}\right] = E\left[-2LL\middle|\mathbf{y}\right] \tag{5.1}$$

where *LL* is the data likelihood given parameter estimates: $LL = \ln\left(P(\mathbf{y}|\boldsymbol{\theta})\right)$. The AIC is then computed as

$$\text{AIC} = \bar{D}(\mathbf{y},\boldsymbol{\theta}) + 2p, \tag{5.2}$$

with *p* denoting the number of freely estimated model parameters in $\boldsymbol{\Theta}$, which also includes the unobserved latent variables at between-level as the current study employs the augmented approach. The BIC is similarly computed as

$$\text{BIC} = \bar{D}(\mathbf{y},\boldsymbol{\theta}) + p\log n, \tag{5.3}$$

where *n* is the total number of observations. The sample size adjusted BIC is computed as

$$\text{saBIC} = \bar{D}(\mathbf{y},\boldsymbol{\theta}) + p\log\frac{n+2}{24}, \tag{5.4}$$

where the sample size *n* in Eq. (5.3) is replaced by (*n* + 2)/24 (Sclove, 1987). An additional information criterion utilized in the current study is the IC1000 developed by Kim and Wang

(2019). It has been found to be more effective in choosing the correct model with large samples. IC1000 is defined as

$$\text{IC1000} = \bar{D}(\mathbf{y}, \boldsymbol{\theta}) + p\frac{n}{1000}, \tag{5.5}$$

which is expected to penalize the model complexity more efficiently under huge log-likelihood.

5.3 Results

With Bayesian estimation, it took 25.8 minutes to fit the 2-class model and 32.5 minutes to fit the 3-class model. Both the Geweke's diagnostic and visual inspection indicate that the estimation for the two models converged after 50,000 iterations. The estimation results are summarized in Table 5-2. The information criteria yield consistent results, collectively suggesting that the 2-class model is a better fit than the 3-class model. Upon a closer examination of the parameter estimates, it can also be noted that the parameter estimates for the first two latent classes obtained under the 3-class model are remarkably similar with the ones obtained under the 2-class model. By forcing a third latent class, however, the 3-class model yields an additional latent class with unstable parameter estimates, presenting challenges in interpretation (e.g., $\hat{\gamma}_{103} = 251.459$, 95% HPDI [-.820, 614.247]). The classification outcomes for both the 2-class and 3-class models are likewise highly similar: the 2-class model classifies 96 schools to class 1 and 66 schools to class 2, while the 3-class model classifies 97 schools to class 1 and 65 to class 2. Although a third latent class is introduced, no schools are assigned to class 3 under the 3-class model. In summary, given the estimation results, the 2-class model should be selected as the final model for further interpretation.

With the 2-class solution, the estimated unconditional probability of a school belonging to latent class 1 is .549. Schools in latent class 1 have an average within-school variability of -

.919 on the log scale, while schools belonging to latent class 2 have an average within-school variability of -.784 on the log scale. The schools in each latent class have varying levels of within-school variability: the inter-school variance of within-school variability is .221 for latent class 1, and is .090 in latent class 2. Therefore, descriptively speaking, while on average the students think more similarly of the teacher support within the same schools in latent class 1, there is greater variability from school to school in terms of the level of homogeneity in latent class 1 compared to class 2. With the mean structure and variance combined, the results suggest that the majority of the schools in latent class 1 have a within-school variability falling in the range of (-1.840, .002), while schools in latent class 2 have a within-school variability in the range of (-1.372, -0.196), all of which are on the log scale.

Within each latent class, the schools also vary in the between-level aggregated perceived teacher support, with an estimated variance of .059 in class 1 and an estimated variance of .047 in class 2. The between-level aggregated perceived teacher support is found to be negatively associated with the within-level dispersion within each latent class—yielding a correlation coefficient of -0.65 in class 1 and -.35 in class 2. It suggests that schools with a higher level of aggregated perceived support is expected to have lower degree of dispersion within school. Additionally, the item intercepts appear to differ between the latent classes. On average, students in latent class 1 report higher level of perceived teacher support than students in latent class 2.

Lastly, with regard to classification decisions, 96 schools are assigned to latent class 1 and 66 are assigned to latent class 2. The mean classification certainty (CC_{mean}) is .867 and the minimum classification certainty (CC_{min}) is .501, suggesting while on average the classification decision is made with relatively high certainty, some schools are classified with greater uncertainty than others—in the worst case scenario, for instance, school 2 could be assigned to

either latent class with nearly equal probability. The model-based class assignment for the randomly sampled 11 schools is displayed in Figure 5-2, with classification certainty ranging from .83 to 1.

To further compare the latent variable approach with the naïve mean score approach, a 2-class naïve mean score model is also fit to the empirical data. The estimation results are presented in Appendix D. It is noted that, with the current example, the naïve mean score model yields distinctively different results than the latent variable model, thus leading to entirely different conclusions. Overall, the naïve mean score model is not able to differentiate between the hypothetical latent sub-populations—the estimated unconditional probability of belonging to one of the latent classes approximates one. According to the classification results, all the 162 schools are assigned to latent class 1. In contrast to the inferences based on the latent variable MMSEM-RV, the naïve mean score model suggests that all the schools are from a single homogeneous population. Compared to the two sub-populations detected by the latent variable model, descriptively, this single population inferred by the naïve model has on average higher within-school dispersion, yet lower between-school variation in the within-level variability. That is to say, on average, the students within each school hold more heterogeneous opinions; but the schools are more similar with one another regarding the dispersion in student perceptions.

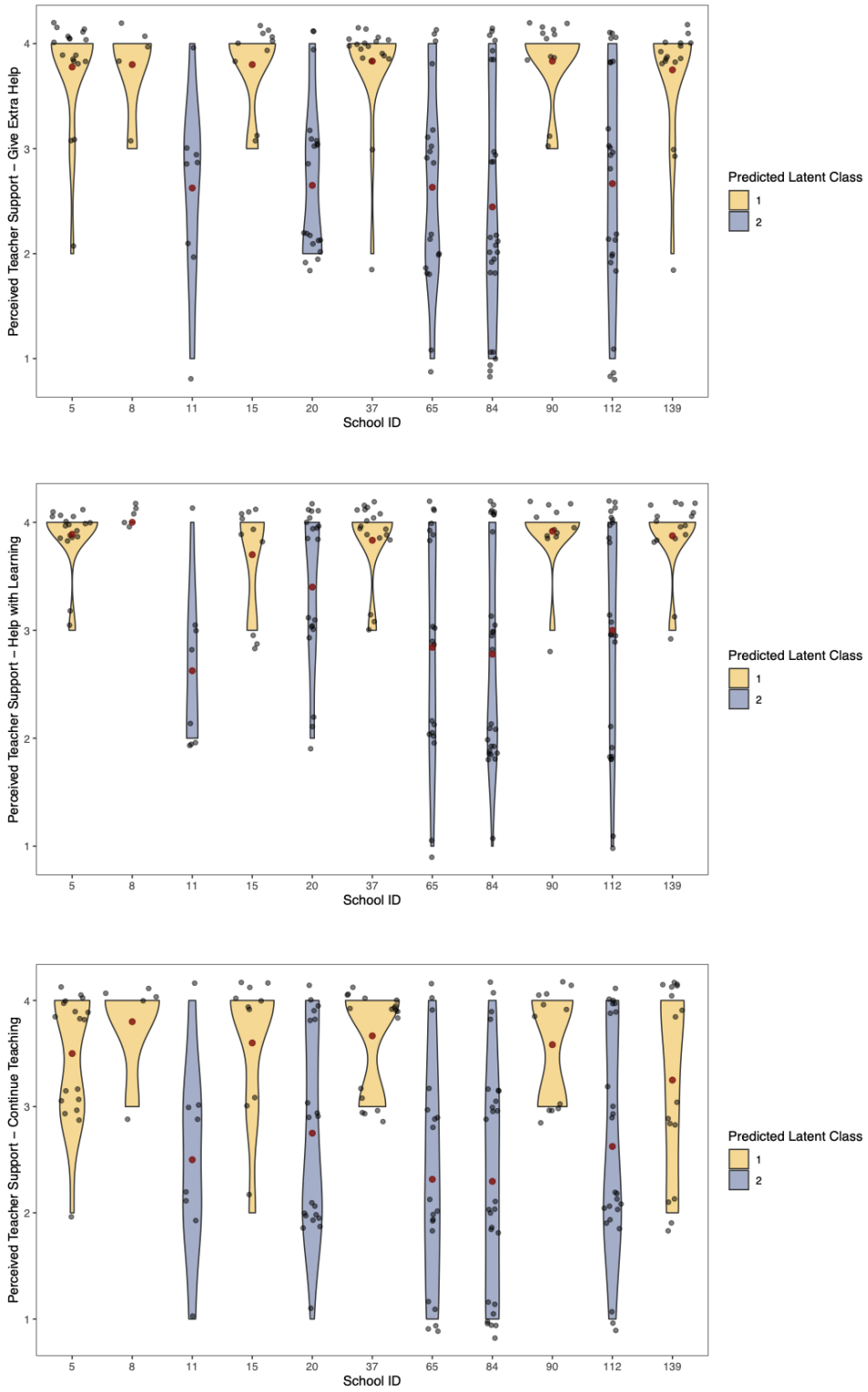
Table 5-2

Estimation Results for the Illustrative Example

		$K = 2$		$K = 3$	
		Posterior Mean	95% HPDI	Posterior Mean	95% HPDI
Focal Parameters	γ_{101}	-.919	(-1.054, -.790)	-.921	(-1.062, -.788)
	γ_{102}	-.784	(-.910, -.649)	-.781	(-.911, -.647)
	γ_{103}	-	-	251.459	(-.820, 614.247)
	π_1	.549	(.397, .702)	.549	(.396, .694)
	π_2	-	-	.445	(.302, .601)
	ϕ_{01}	.059	(.036, .085)	.059	(.035, .084)
	ϕ_{101}	-.074	(-.117, -.034)	-.073	(-.117, -.034)
	ϕ_{11}	.221	(.110, .346)	.220	(.110, .349)
	ϕ_{02}	.047	(.028, .069)	.047	(.029, .069)
	ϕ_{102}	-.023	(-.051, .001)	-.023	(-.051, .002)
	ϕ_{12}	.090	(.042, .146)	.090	(.043, .148)
	ϕ_{03}	-	-	.338	(.045, .893)
	ϕ_{103}	-	-	.0001	(-.397, .428)
	ϕ_{13}	-	-	.334	(.046, .882)
Peripheral Parameters	α_{11}	3.376	(3.304, 3.452)	3.375	(3.303, 3.444)
	α_{21}	3.458	(3.392, 3.527)	3.458	(3.388, 3.522)
	α_{31}	3.099	(3.023, 3.178)	3.097	(3.022, 3.172)
	α_{12}	3.113	(3.024, 3.201)	3.112	(3.030, 3.194)
	α_{22}	3.307	(3.218, 3.386)	3.306	(3.228, 3.383)
	α_{32}	2.885	(2.793, 2.970)	2.885	(2.801, 2.965)
	α_{13}	-	-	2.576	(-3.531, 8.911)
	α_{23}	-	-	2.602	(-3.442, 8.882)
	α_{33}	-	-	2.609	(-3.514, 8.933)
	λ_2	1.010	(.971, 1.050)	1.009	(.970, 1.050)
	λ_3	1.014	(.969, 1.058)	1.013	(.969, 1.058)
	$\Theta_{11}^{(W)}$	.162	(.134, .190)	.163	(.136, .189)
	$\Theta_{21}^{(W)}$	.097	(.078, .116)	.097	(.078, .115)
	$\Theta_{31}^{(W)}$	.388	(.348, .427)	.388	(.349, .426)
	$\Theta_{12}^{(W)}$	.323	(.281, .365)	.322	(.280, .364)
	$\Theta_{22}^{(W)}$	.184	(.151, .220)	.186	(.151, .221)
	$\Theta_{32}^{(W)}$	.464	(.418, .510)	.464	(.418, .510)
	$\Theta_{13}^{(W)}$	-	-	.252	(.046, .620)
	$\Theta_{23}^{(W)}$	-	-	.250	(.043, .611)
	$\Theta_{33}^{(W)}$	-	-	.247	(.042, .601)
Information Criteria	Deviance	13306.903		13311.373	
	AIC	14648.903		15323.373	
	BIC	18729.157		21440.715	
	saBIC	16597.098		18244.215	
	IC1000	15475.575		16562.765	

Figure 5-2

Latent Class Assignment of a Random Sample of 11 Schools



Chapter 6. Discussion

6.1 Conclusion

Variability is an important topic in many research and applied settings. Inferences about intraindividual variability (i.e., consistency of an individual over multiple measurements) as well as intragroup variability (i.e., cohesiveness among members within a group or team) are often of key interest across a variety of research scenarios (e.g. Adolph et al., 2015; Cox & Blake, 1991; Easley, 2001; Fagot et al., 2018; Hultsch & MacDonald, 2004; Jehn et al., 1999; Kupers et al., 2019; Petersen et al., 2004; van Knippenberg et al., 2004). Given that variability carries important information about human development and organizational behavior, it is thus of critical importance to identify unobserved subpopulations that differ in the patterns of variability. For instance, some individuals may share a similar predisposition of being highly unstable in their mental status, which renders them more susceptible for certain mental disorders; likewise, some organizations may collectively demonstrate an exceptionally higher degree of cohesiveness relative to other organizations, which may have implications for future market performance.

As the theoretical importance of variability becomes increasingly recognized, different methodological approaches have been introduced to study variability as a random variable. The available methods include more intuitive multi-stage approaches involving simple summary statistics (e.g., SD), as well as more sophisticated statistical models such as MELSM (Hedeker et al., 2008) and MSEM-RV (Feng & Hancock, 2022). These methodological advancements offer useful analytical tools for researchers to model variability as the focal construct. Nonetheless, in comparison to mean modeling, variability modeling is still a relatively new addition to the methodological literature. Particularly, a huge challenge still remains when it comes to detecting the unobserved heterogeneity in random variability. Given that no existing research addresses the

methodological considerations on this issue, the current study aims to develop a model-based approach that can be employed to effectively identify the underlying groups based on the distinct patterns of random variability.

The current study proposed an innovative latent variable based modeling framework, which specifically aims to detect unobserved sub-populations that vary in random intraindividual or intragroup variability. Building upon the previous work in MSEM and MSEM-RV, the theoretical framework of MMSEM-RV was articulated in the current study. Compared to MLM based approaches, it capitalizes on the strengths of SEM, such that measurement error can be properly accommodated by explicitly incorporating a measurement model. As a finite mixture extension of MSEM-RV, MMSEM-RV offers greater flexibility in modeling random variability as it allows for both continuous random effects and latent categorical variables in explaining the heterogeneity in within-level variability.

Given the structural complexity of MMSEM-RV, Bayesian inference via MCMC was proposed for parameter estimation. The performance of MMSEM-RV with Bayesian estimation was systematically evaluated through simulation studies. The practical application of MMSEM-RV was also demonstrated using a real-world data example. As shown in both the simulation study and the empirical data analysis, Bayesian estimation via Gibbs sampling is an effective approach for estimating MMSEM-RV, although it may require longer chains of iterations and better-informed priors to facilitate estimation convergence when there is a large sample size or when the model grows in complexity.

The simulation study suggests that MMSEM-RV have consistently satisfactory classification performance across all the simulated conditions. That is to say, MMSEM-RV is an effective parametric approach for identifying sub-groups of individuals characterized by distinct

patterns of intraindividual stability, or sub-groups of clusters that have distinct patterns of intragroup cohesiveness.

With regard to specific model parameters, MMSEM-RV yields accurate or approximately accurate estimates for the mixing proportion and between-level fixed effects. In most conditions, MMSEM-RV can be employed to effectively infer about the proportion of clusters belonging to each latent class, as well as the average within-level variability for each latent class. In contrast, however, it is most challenging to estimate the between-level random effects variance and covariance. According to the simulation results, MMSEM-RV tends to overestimate the between-level variance of within-level variability as well as its covariance with the between-level latent construct. This upward bias gets worsened with smaller sample sizes. This is not unexpected as the MLM literature has reported the evident estimation bias with random effect variance in small samples (Hoyle & Gottfredson, 2015; D. M. McNeish & Stapleton, 2016). In the context of MMSEM-RV, the need for a large sample size may become even more crucial to ensure accurate results. This is because we are essentially trying to estimate the variance of variance—a second-order variance parameter. Generally speaking, the simulation results suggest that a large sample size (particularly a large number of clusters) can effectively reduce the estimation bias in all the focal parameters. But to obtain unbiased estimation of the between-level variance and covariance for random variability, a substantially larger sample size than what the current study considered may be required. In light of this, caution needs to be exercised when interpreting the estimates of and drawing inferences about the between-level variance and covariance of variability.

In addition to sample size, the simulation results also revealed the important impact of intercept separation on estimation accuracy of MMSEM-RV. All else being equal, a wider class

separation in item intercepts can considerably reduce estimation bias. This is consistent with previous literature on mixture models (e.g., Depaoli, 2013; Tofighi & Enders, 2008; Tueller & Lubke, 2010), which has documented that a poor class separation can greatly compromise the parameter estimation and correct class enumeration. Intriguingly, the simulation results further suggest that the estimation accuracy is particularly problematic when small intercept separation is accompanied with observed measures that are of good composite reliability. Although it may seem counterintuitive on a first thought—how could having a good measure be a bad thing?—this finding is, in fact, well in line with earlier findings in the mixture models literature. Because given all else equal, the class separation actually decreases as the measurement quality improves. To better elaborate on this point, we can consider the model implied "average" variance of the observed indicators in latent class *k*, which can be computed as

$$\boldsymbol{\Sigma}\left(\mathbf{Y}\middle|C_j = k, \boldsymbol{\theta}\right) = \boldsymbol{\Lambda}_k^{(\mathrm{B})}\phi_{0k}\boldsymbol{\Lambda}_k^{(\mathrm{B})\prime} + \boldsymbol{\Theta}^{(\mathrm{B})} + \boldsymbol{\Lambda}_k^{(\mathrm{W})}\exp(\gamma_{10k})\boldsymbol{\Lambda}_k^{(\mathrm{W})\prime} + \boldsymbol{\Theta}^{(\mathrm{W})}. \tag{6.1}$$

It can easily tell that with the same between-level fixed effects, random effect variance, and residual variance, the total variance of **Y** goes up as the factor loadings go up. An inflated total variance in **Y**, in turn, has important implications for class separation. If Mahalanobis distance (MD) is used to measure the class separation with regard to the observed measures

$$\mathrm{MD} = \sqrt{(\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2)'\boldsymbol{\Sigma}_{pooled}^{-1}\left(\mathbf{Y}\middle|\boldsymbol{\theta}\right)(\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2)}, \tag{6.2}$$

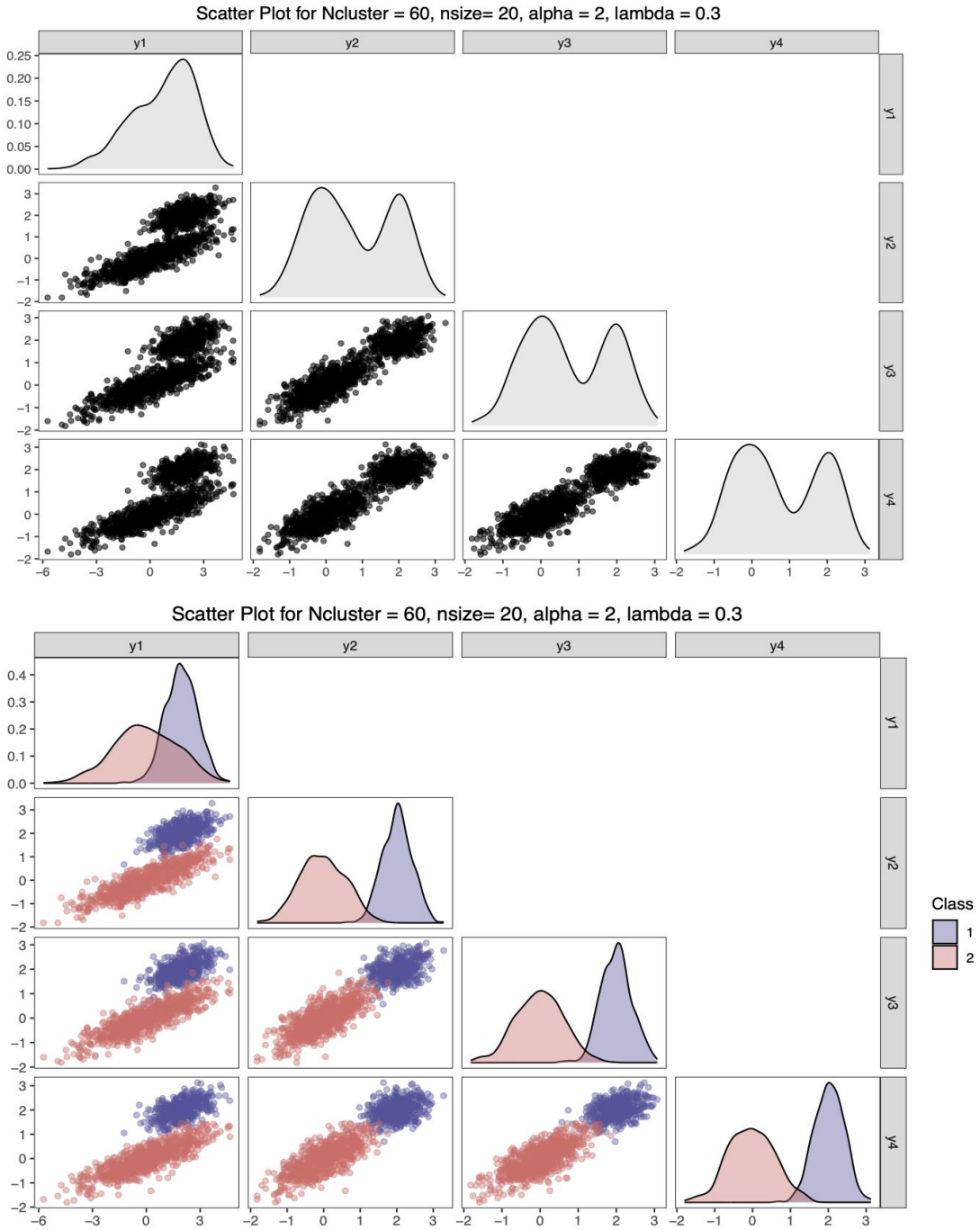
it can be shown that given the same intercept separation, MD would decrease as the total variance in **Y** increases. Therefore, having a good observed measure is not always beneficial for the sake of mixture modeling. To illustrate this point in a more intuitive way, the distributional plots of the simulated data generated for the population analysis under two different conditions are displayed in Figure 6-1 through Figure 6-4. As evident in the plots, comparing the condition where $\alpha_1 = 2$ and $\lambda = 0.3$ against the condition where $\alpha_1 = 2$ and $\lambda = 0.8$, there is a greater degree

of overlapping between the two classes in the latter condition. Notably, with a high factor loading, the marginal distribution of the observed indicators from the two latent classes almost perfectly overlap with each other (Figure 6-3), thereby bringing challenges for the model to correctly distinguish between the latent classes and estimating the class-specific parameters.

Although having a good measure is not necessary for satisfactory model performance, what can be way more detrimental is to ignore the measurement error altogether. The simulation results also highlight the importance of properly taking measurement error into consideration for the purpose of modeling unobserved sub-populations with distinct patterns of random variability. Treating the observed measures as if they were free of measurement error, the naïve mean score model yields larger estimation bias, worse interval coverage, and less precision in parameter estimates compared to its latent construct counterpart. The results obtained from the naïve approach can be particularly misleading if researchers are interested in making inferences about class-specific parameters. The only exception to this may be in a case when researchers are only interested in making classification. The treatment of measurement error does not seem to impact the classification accuracy under the various conditions considered in the current study.

Figure 6-1

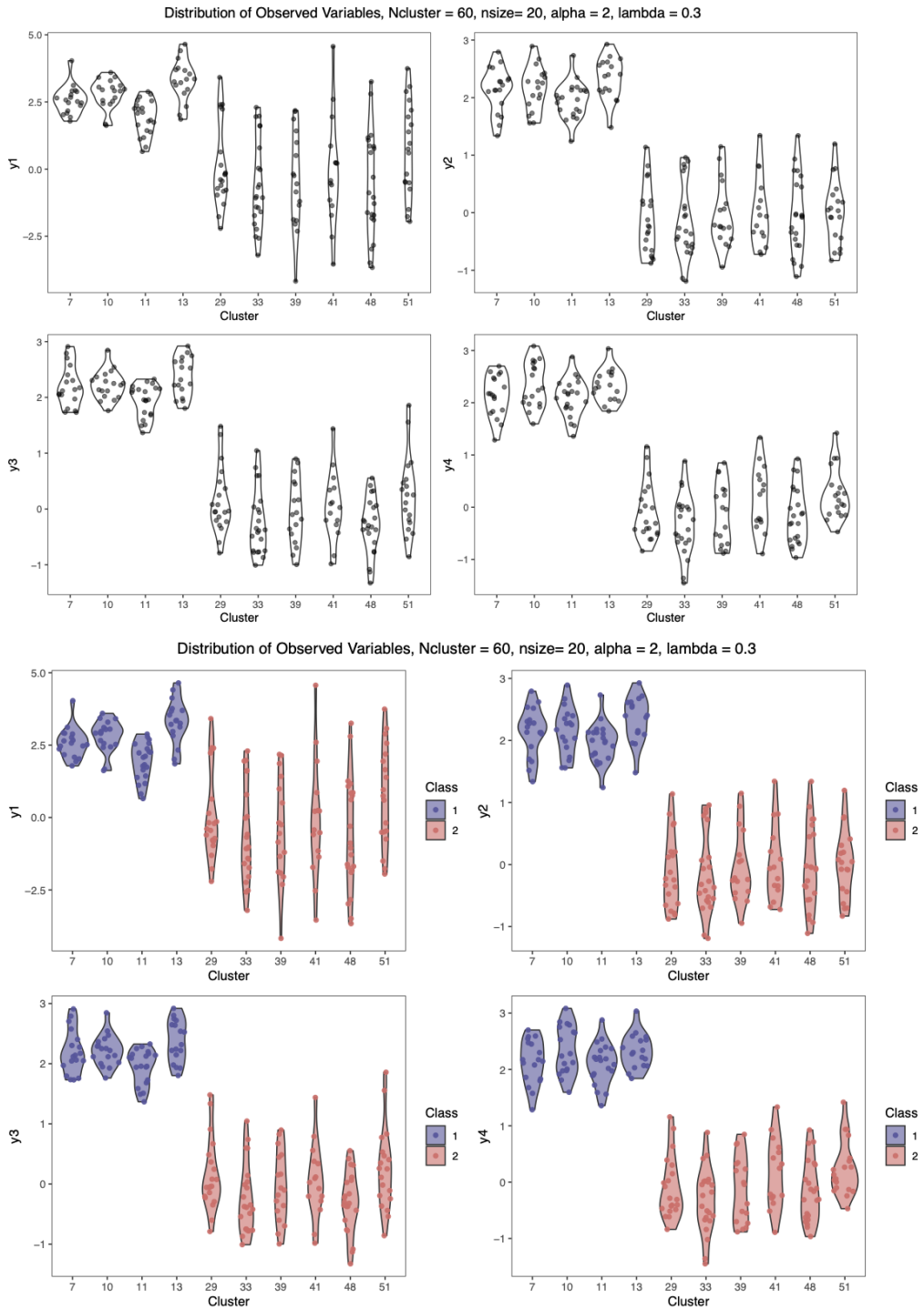
Scatter Plots for a Simulated Data with Small Intercept Separation and Low Factor Loading



Note. The lower panel shows the actual class membership in data generating process.

Figure 6-2

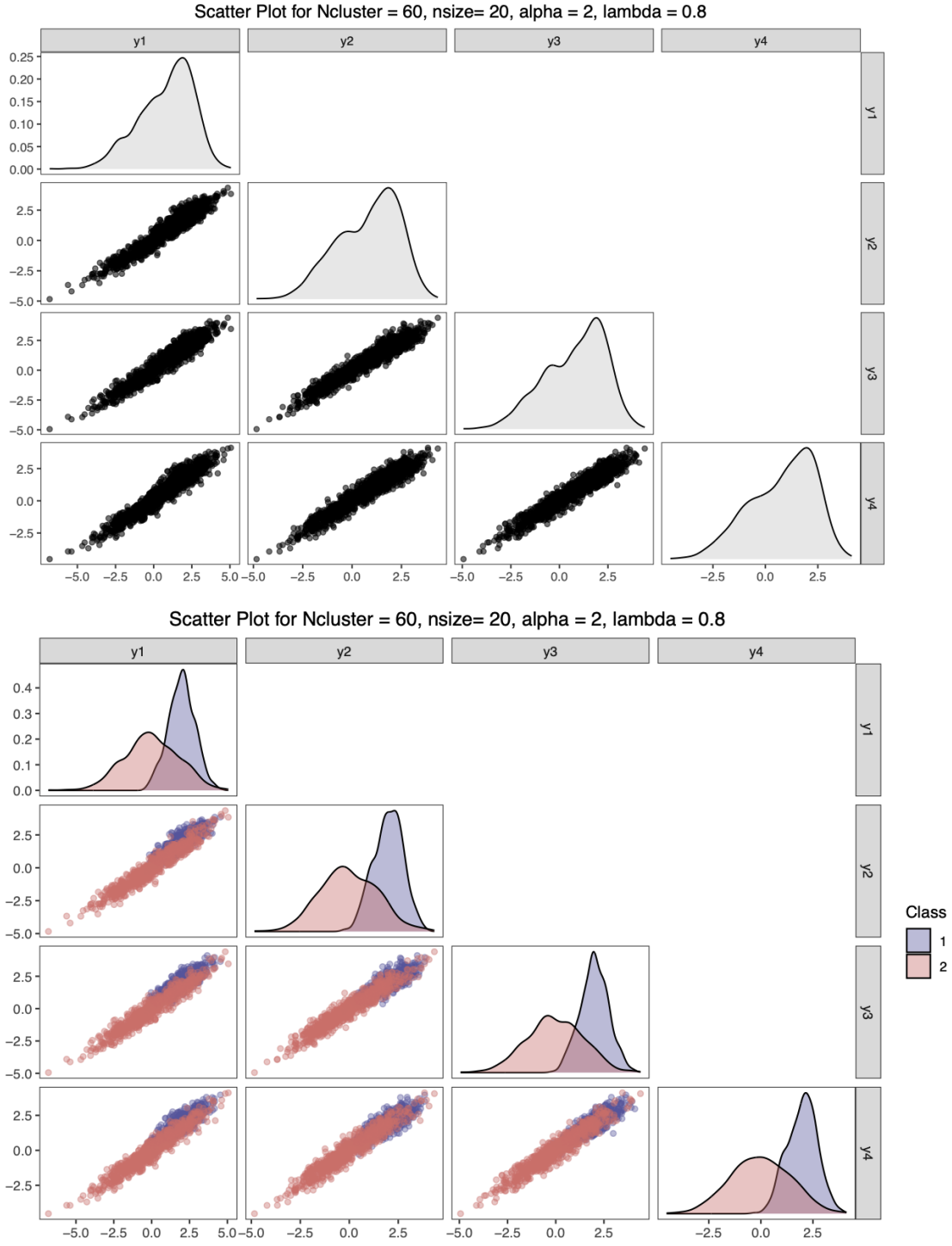
Violin Plots for a Random Sample of 10 Schools in a Simulated Data with Small Intercept Separation and Low Factor Loading



Note. The lower panel shows the actual class membership in data generating process.

Figure 6-3

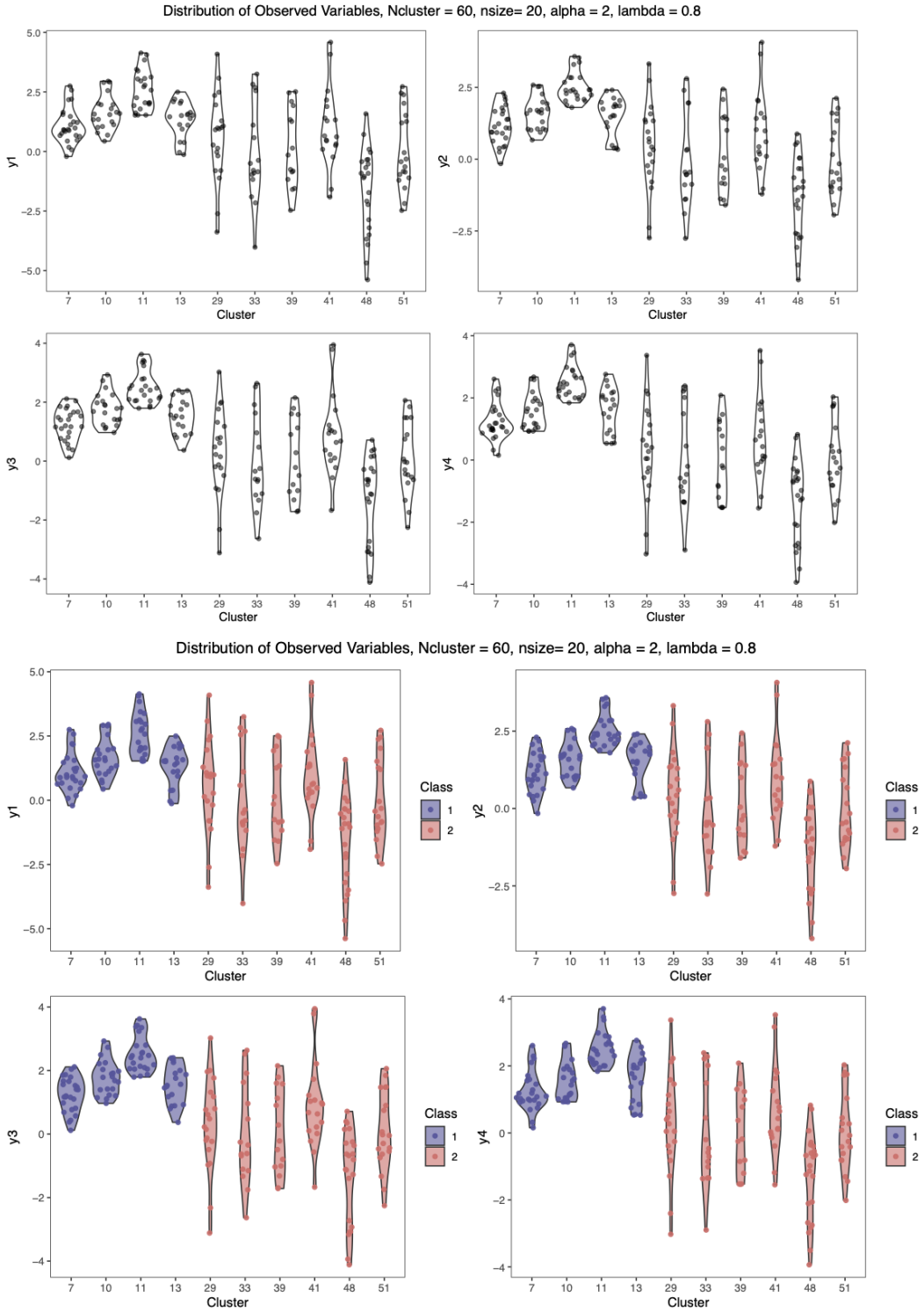
Scatter Plots for a Simulated Data with Small Intercept Separation and High Factor Loading



Note. The lower panel shows the actual class membership in data generating process.

Figure 6-4

Violin Plots for a Random Sample of 10 Schools in a Simulated Data with Small Intercept Separation and High Factor Loading



Note. The lower panel shows the actual class membership in data generating process.

6.2 Limitations and Future Directions

While the current study is undoubtedly an important addition to the field of random variability modeling, which positively contributes to both the methodological theory and applied research, there are some caveats that need to be noted as well. Acknowledging these limitations will help inform future research directions. Firstly, due to limited time and computational resources, the current study did not implement Monte Carlo simulation for every manipulated condition. As a result, the simulation study was not able to fully examine the behavior of Bayesian estimator for MMSEM-RV over repeated sampling across various conditions. The population analysis provided useful preliminary results regarding the model performance. However, population analysis was also subject to estimation error and thus limited the interpretation when extreme results were obtained. Future research will extend the current simulation study by conducting Monte Carlo simulation under the full-factorial design. Additionally, in order to assess the sample size requirement for obtaining accurate estimates of the between-level random effects variance, future study can increase the number of clusters in an expanded simulation design than what was considered in the current study.

Secondly, another important issue to be addressed in future research is to understand the implications of violating model assumptions and how it impacts classification, parameter estimation, and statistical inference for MMSEM-RV. For the current study, it is assumed that all the model assumptions are met throughout both the simulation study and the empirical illustrative example[9]. Therefore, it remains unclear how potential departures from these

[9] The illustrative example assumes measurement invariance, no omitted covariates, and multivariate normality, which can be potentially violated in the empirical data. For example, the resources allocated to different schools and school districts can vary vastly (e.g., professional development training opportunities for teachers, classroom equipment, class sizes, etc.). Those factors were not measured for the current analytical sample and thus were not controlled for in the analysis. However, they can potentially induce spurious clustering of student perceived support as provided by their teachers. As

assumptions may adversely affect the model performance. It would be of great value for future research to assess the robustness of Bayesian estimation for MMSEM-RV in face of assumption violation. Some assumptions may be relaxed without compromising the estimation accuracy or efficiency, while violating other assumptions may lead to more seriously undesirable consequences. It is also an important avenue for future research to develop a principled approach for assumption checking. Importantly, it will be helpful to provide multiple empirical examples across different research scenarios, demonstrating how to properly assess the model assumptions in practice for the applied researchers.

Thirdly, although it is outside of the scope of the current study to develop a formal model fit assessment procedure for MMSEM-RV, a thorough investigation into model fit assessment within the context of MMSEM-RV in future research would be of great value. Given the inherent complexity in the model structure, the adverse impact of model misspecification at any layer can propagate across the entire model. For example, any misspecification in the measurement model (e.g., measurement noninvariance) can potentially lead to incorrect inferences about latent classes. Therefore, it is critical to have an effective approach to detect model misspecification in MMSEM-RV. Particularly, it would be of great importance to isolate the model misfit with the measurement model from the misfit with the structural model of random variability. Although the existing literature provided useful guidance on model fit for MSEM (Ryu, 2014; Ryu & West, 2009) and Bayesian SEM (Garnier-Villarreal & Jorgensen, 2020; Levy, 2011), the knowledge about model fit assessment for MSEM-RV, MMSEM-RV, or random variability modeling in general is seriously lacking. Without a principled guidance, the current study was only able to briefly demonstrate the use of information criteria for model comparison, whose

discussed in chapter 6, future research is needed to develop a principled way of assumption checking and to investigate the robustness of MMSEM-RV to assumption violations.

performance was not systematically evaluated. Therefore, it will be beneficial if future methodological research can investigate the effective procedures and develop appropriate fit indices for model fit assessment of random variability models. Additionally, the systematic evaluation of the performance of these fit indices through rigorous simulation studies would provide valuable insights for researchers and practitioners.

A related consideration concerns class enumeration in MMSEM-RV. While the specification of the measurement model and the structural model for latent random variability should be informed and guided by theoretical considerations, the process of determining the optimal number of latent classes is fundamentally exploratory. Prior to fitting the mixture model, examining the descriptive plots, as shown in Chapter 1 and Chapter 5, can provide helpful directions. The currently study also demonstrated the use of information criteria for the purpose of class enumeration in the illustrative example. Additionally, although not articulated in the study, descriptive statistics are also useful for preliminary assessment of population heterogeneity in random variability. For example, for intraindividual variability, descriptive measures such as iSD, iMSSD, iSDr, CV, or range of the observed outcome can be calculated and examined to identify any potential sub-groupings prior to model fitting. Given the data-driven nature of class enumeration, however, it is crucial to understand how misfit in the confirmatory part of the model may influence this exploratory process. Specifically, it is important to investigate whether different combinations of confirmatory model specifications and latent class enumerations yield equivalent model fit, and if so, how to better determine the number of latent classes when there exists model misfit in the measurement model or structural model.

Fourthly, the current study did not examine how sensitive the estimation of MMSEM-RV is to different priors and starting values. In the simulation study and empirical data analysis, informative priors were employed for the item intercepts to facilitate better model convergence. But it is not yet clear about how stable the estimation results are if different priors were specified, and whether any misalignment between the priors and the population parameter values will result in biased results. Regarding the starting values, for the simulation study, starting values are carefully chosen based on the known population parameter values. For any empirical data analyses, the population values of model parameters are generally unknown and thus it is more challenging to specify appropriate starting values. Will random starting values work equally well as the more informed starting values? These remaining questions warrant further investigation in future research.

Lastly, the current study only considered and evaluated the results for the unconditional MMSEM-RV model, where no structural relation is specified at any level. However, it can be of great interest to employ conditional MMSEM-RV in many research scenarios. For instance, researchers may be interested in exploring whether the relation between group cohesiveness and group-level performance differs among unobserved sub-populations. As another example, researchers may be curious about whether there are unobserved sub-population of individuals, for whom the intraindividual consistency plays a different role in predicting future mental health. In these cases, conditional MMSEM-RV can be a useful analytical tool. Therefore, future research could extend the current study and look into the performance of conditional MMSEM-RV under varying conditions.

With all being said, the current study on MMSEM-RV serves an important addition and unique contribution to the literature, given that no existing research yet addresses the

methodological considerations of modeling unobserved sub-populations with differing random variability. It has a broad appeal across various fields for applied research that studies random intraindividual variability over repeated measures or random intragroup variability as the focal construct of inferential interest. With the current study, a solid foundation has been laid for further advancements in the understanding and application of MMSEM-RV, setting up the stage for many more valuable contributions to come.

Appendix A. Common Measures of Intraindividual Variability

The iSD for each individual *i* is computed as

$$iSD_i = \sqrt{\frac{1}{T_i - 1}\sum_{t=1}^{Ti}(y_{i(t)} - \overline{y_i})^2}\ , \tag{A.1}$$

where $y_{i(t)}$ is the outcome variable for individual *i* at time *t*, and $\bar{y}_i$ is the individual average across all the T_i measurements.

The iMSSD for each individual *i* is computed as

$$iMSSD_i = \frac{1}{T_i - 1}\sum_{t=2}^{T_i}(y_{i(t)} - y_{i(t-1)})^2\ . \tag{A.2}$$

The iSDr for each individual *i* is computed as

$$iSDr_i = \sqrt{\frac{1}{T_i - 1}\sum_{t=1}^{Ti}(y_{i(t)} - y_{i(t)})^2}\ , \tag{A.3}$$

where $y_{i(t)}$ is the predicted outcome value for individual *i* at time *t*, which is usually obtained from regression analysis that controls for any systematic within- and between-individual variation in the outcome.

The CV is computed as

$$CV_i = \frac{iSD_i}{\overline{y_i}}\ . \tag{A.4}$$

Appendix B. Simulation Study Results

Table B1

MMSEM-RV Model Convergence and Correct Classification Rate across the Simulated Conditions for the Population Analysis

				Latent MMSEM-RV		Naïve MMSEM-RV	
J	n_j	α	λ	Model Convergence	Correct Classification Rate	Model Convergence	Correct Classification Rate
60	20	2	0.3	1	1	1	0.98
60	20	2	0.8	1	0.95	1	1
60	20	8	0.3	1	1	1	1
60	20	8	0.8	1	1	1	1
60	80	2	0.3	0	1	1	0.98
60	80	2	0.8	1	0.98	1	0.97
60	80	8	0.3	1	1	1	1
60	80	8	0.8	1	1	0	1
100	20	2	0.3	1	1	1	0.91
100	20	2	0.8	1	1	1	0.92
100	20	8	0.3	1	1	1	1
100	20	8	0.8	1	1	1	1
100	80	2	0.3	0	1	1	0.89
100	80	2	0.8	1	0.94	1	0.89
100	80	8	0.3	1	1	1	1
100	80	8	0.8	0	1	1	1
300	20	2	0.3	1	1	1	0.93
300	20	2	0.8	1	0.99	0	0.91
300	20	8	0.3	1	1	1	1
300	20	8	0.8	1	1	1	1
300	80	2	0.3	1	1	1	0.94
300	80	2	0.8	1	0.98	1	0.92
300	80	8	0.3	0	1	1	0.99
300	80	8	0.8	0	1	1	1
600	20	2	0.3	1	1	1	0.94
600	20	2	0.8	1	0.97	0	0.94
600	20	8	0.3	1	1	1	1
600	20	8	0.8	1	1	1	1
600	80	2	0.3	1	1	1	0.92
600	80	2	0.8	1	0.99	1	0.93
600	80	8	0.3	1	1	1	1
600	80	8	0.8	1	1	1	1

Note. Conditions that do not reach estimation convergence are marked in grey.

Figure B1

Estimation Relative Bias for Within-level Residual Variances $\boldsymbol{\Theta}^{(W)}$ *in Latent Class 1*

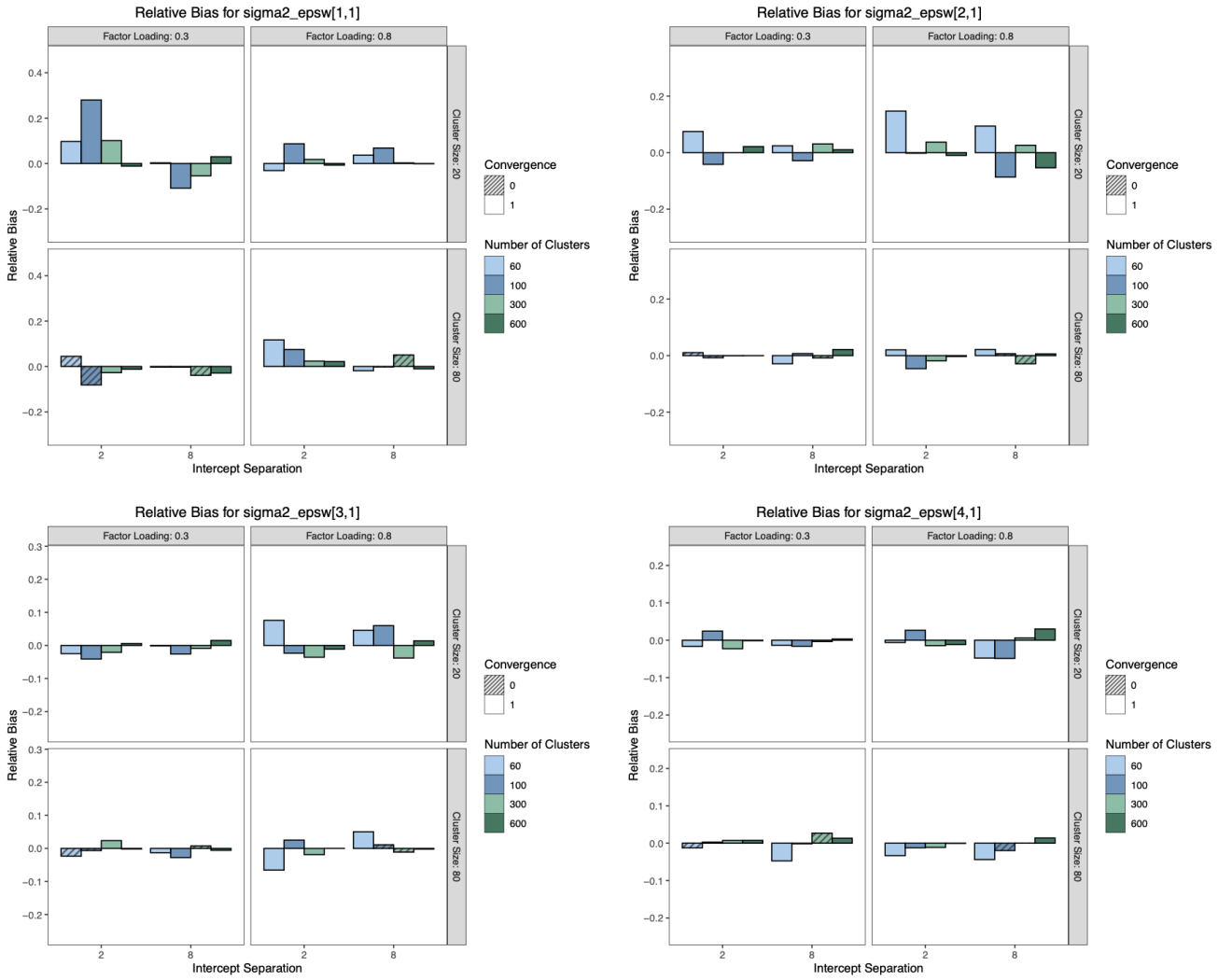


Figure B2

Estimation Relative Bias for Within-level Residual Variances $\boldsymbol{\Theta}^{(W)}$ *in Latent Class 2*

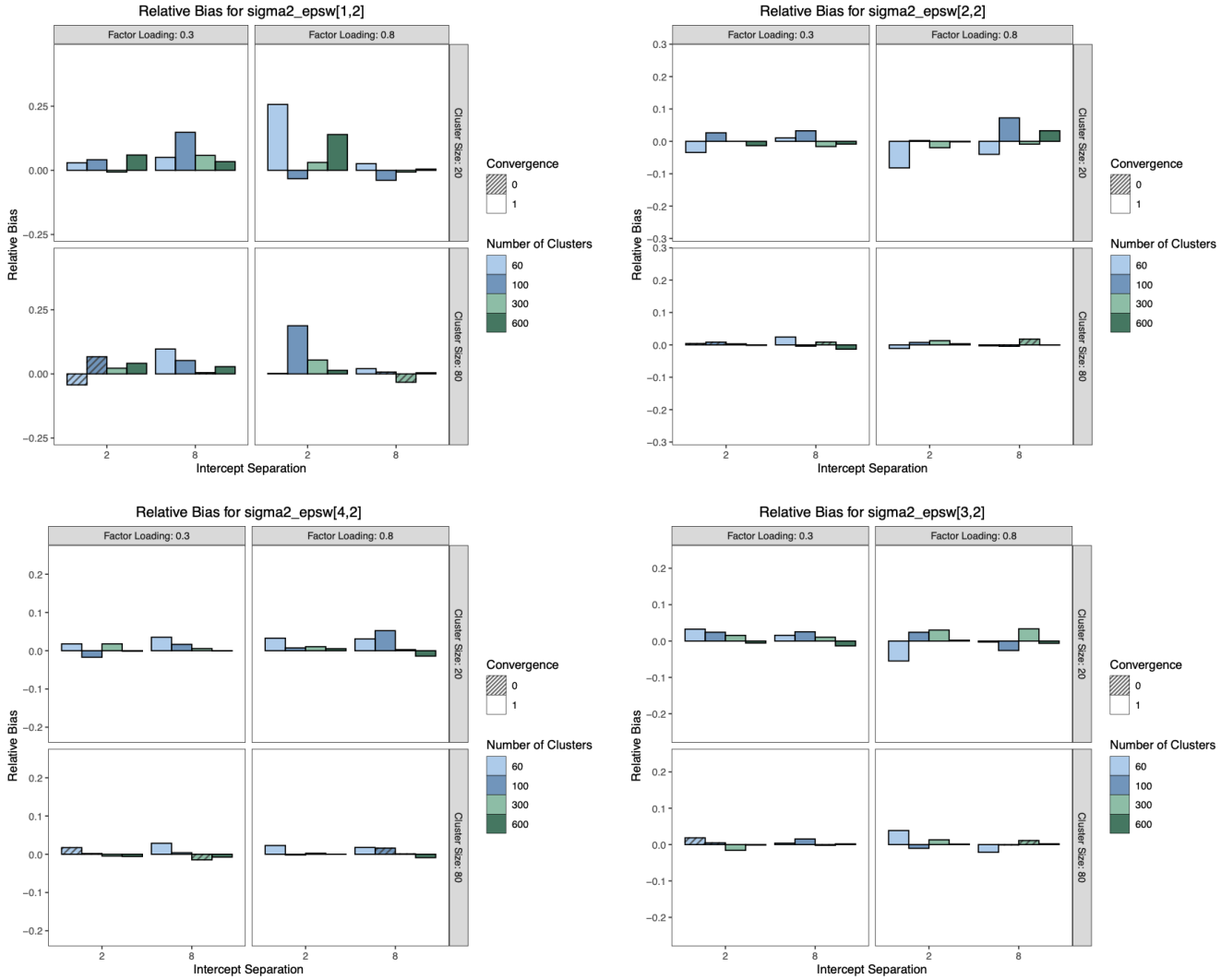


Figure B3

Estimation Relative Bias for Factor Loadings **Λ**

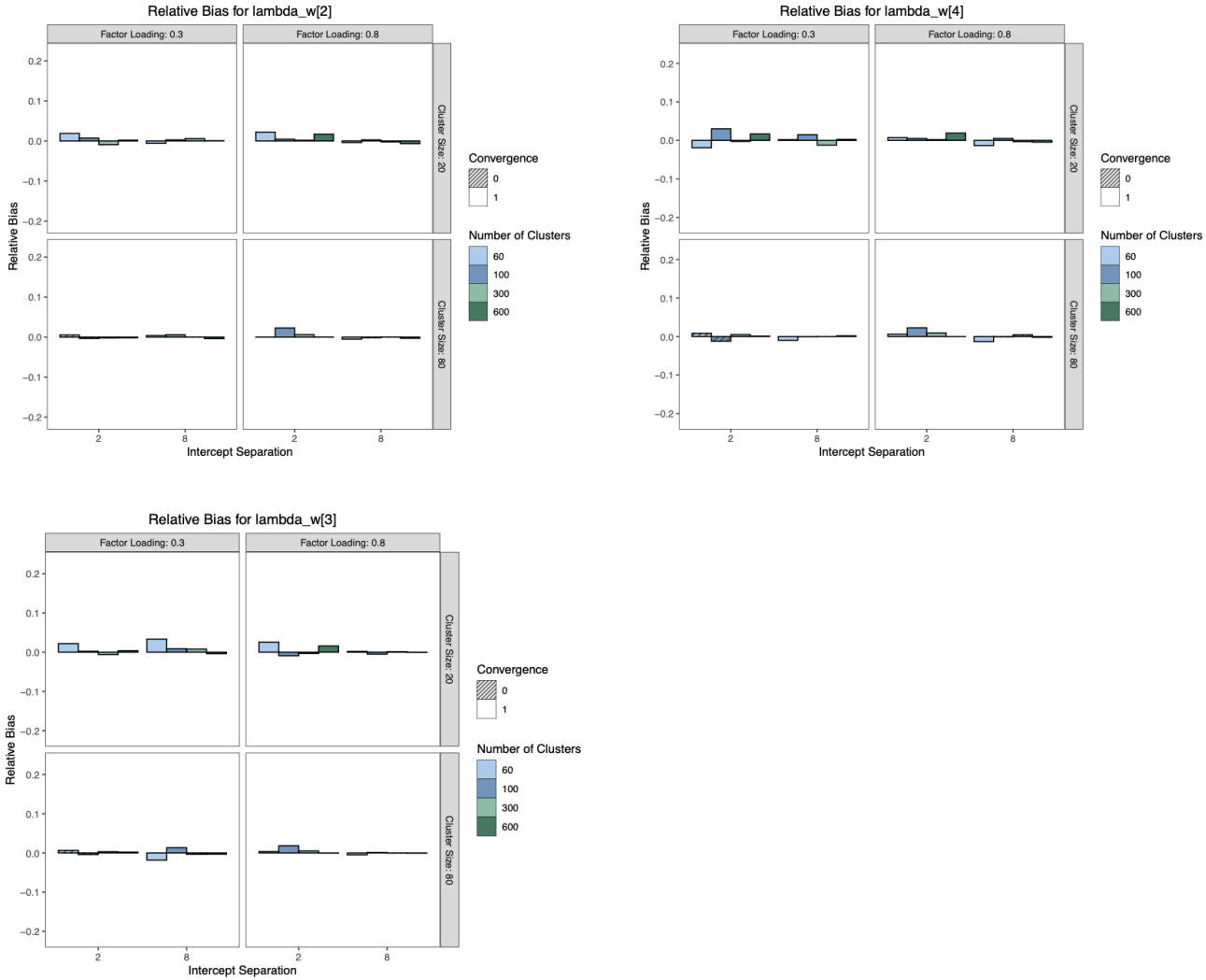


Figure B4

Estimation Relative Bias for Item Intercepts **α** *in Class 1*

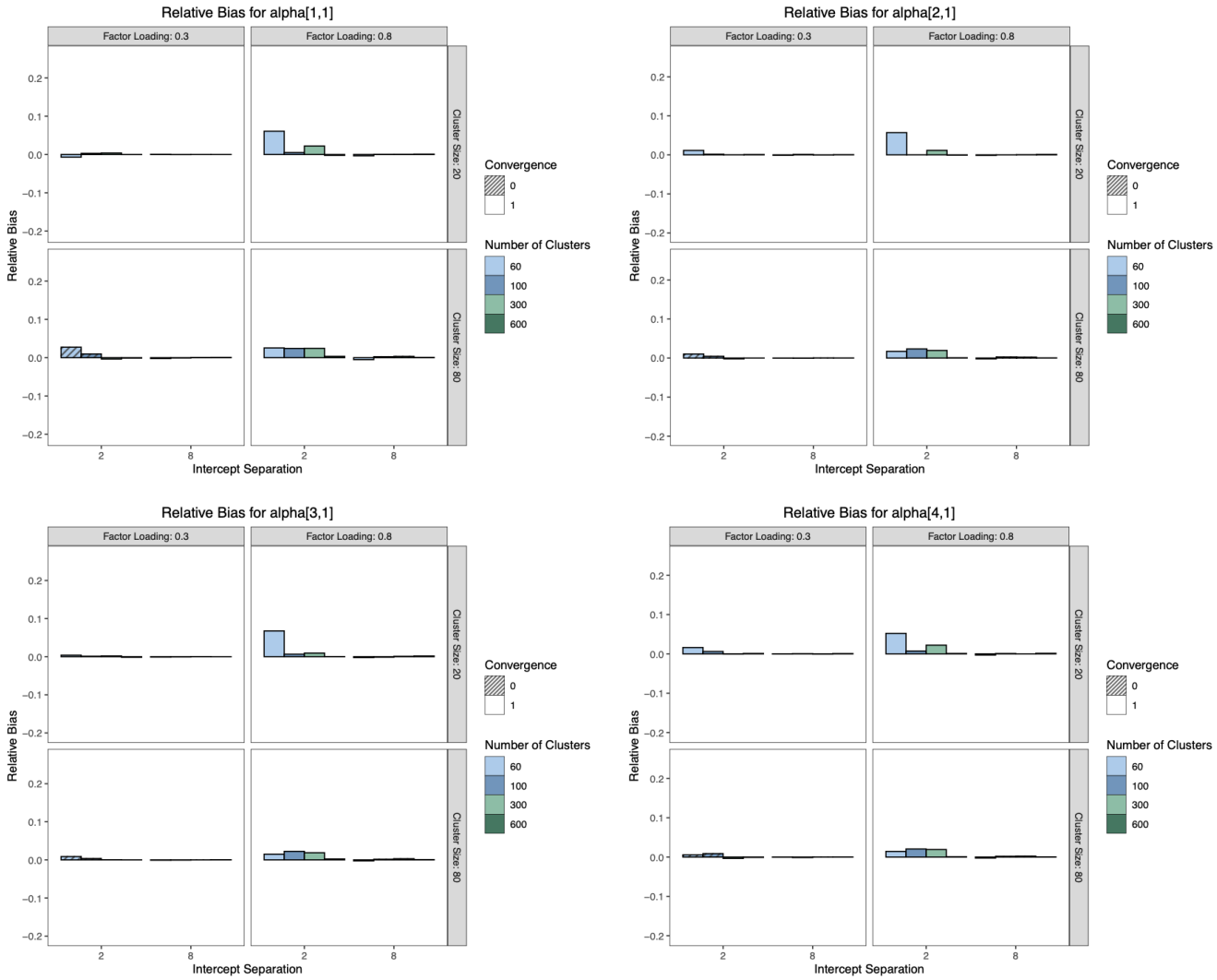


Figure B5

Estimation Relative Bias for Item Intercepts **α** *in Class 2*

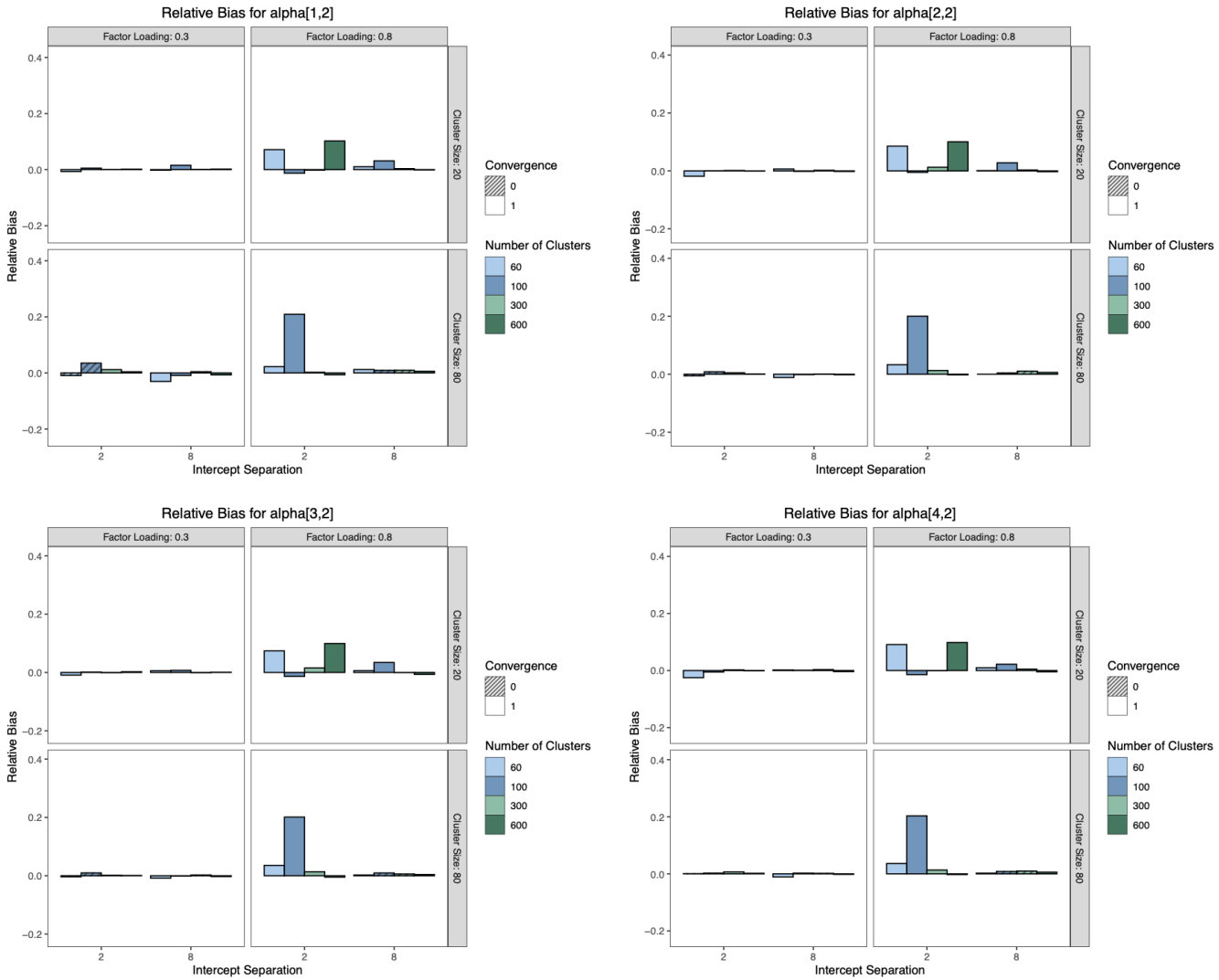


Figure B6

95% HPDI Coverage for Estimating Within-level Residual Variances $\boldsymbol{\Theta}^{(W)}$ *in Latent Class 1*

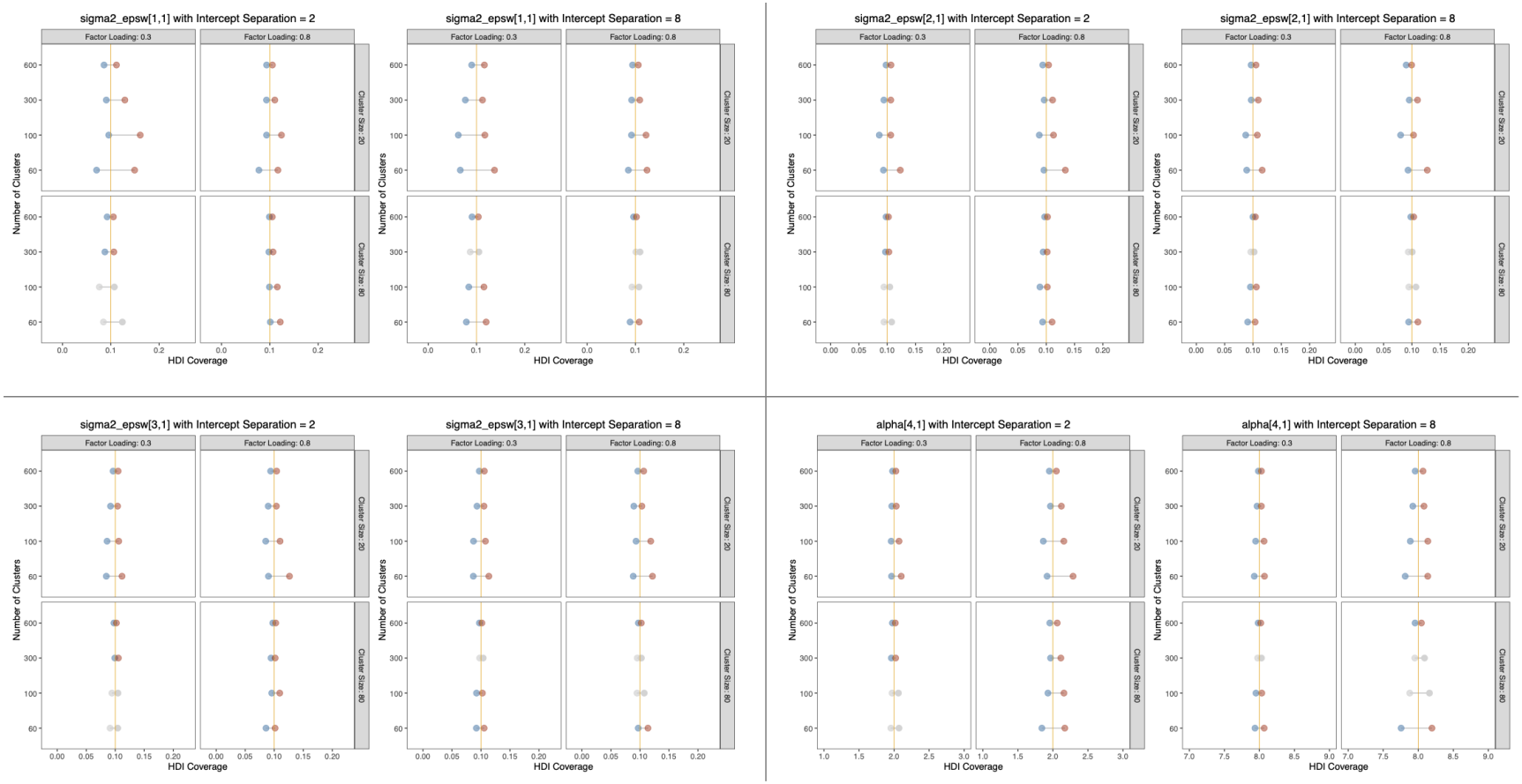


Figure B7

95% HPDI Coverage for Within-level Residual Variances $\boldsymbol{\Theta}^{(W)}$ *in Latent Class 2*

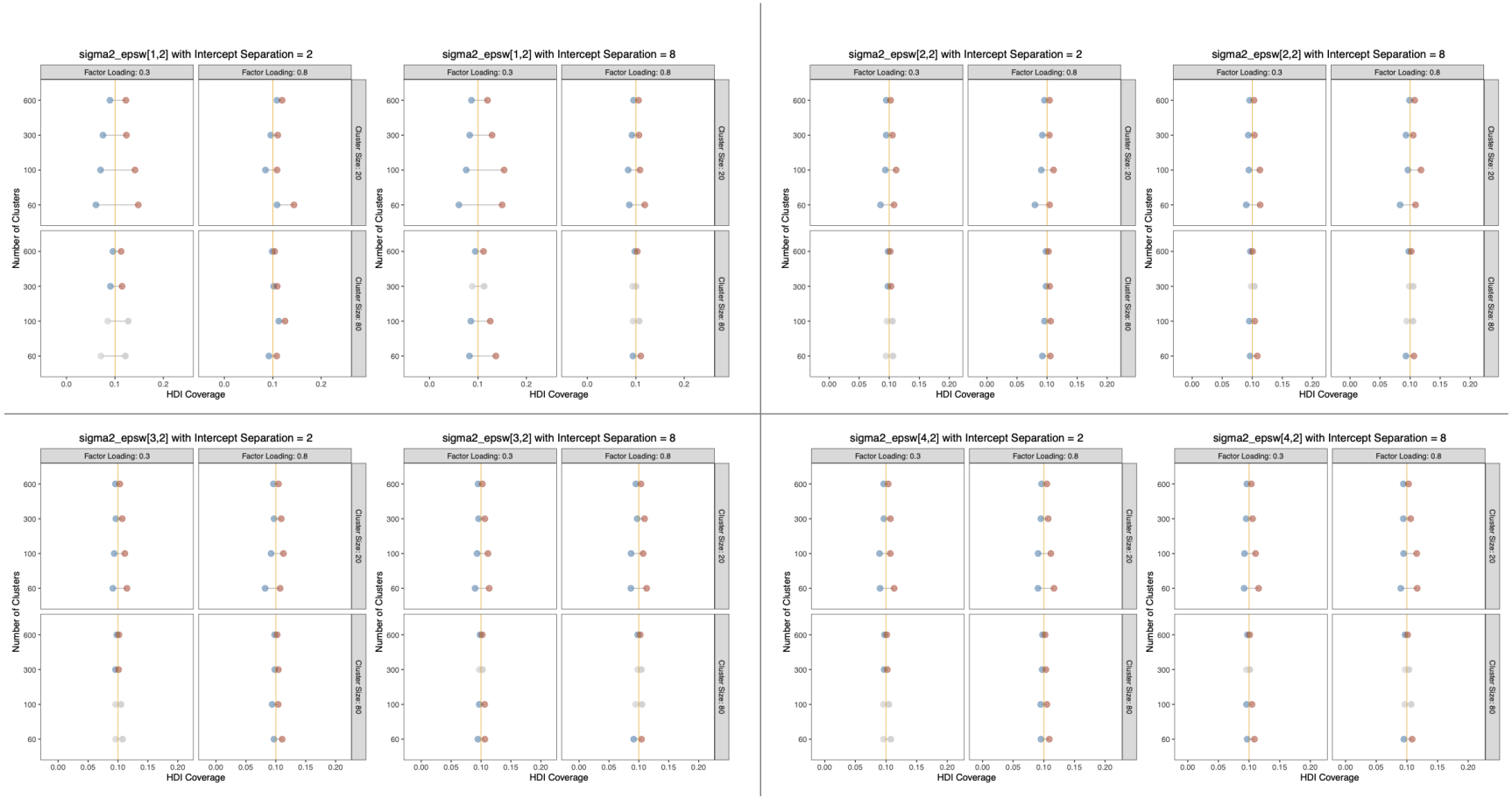


Figure B8

95% HPDI Coverage for Factor Loadings **Λ**

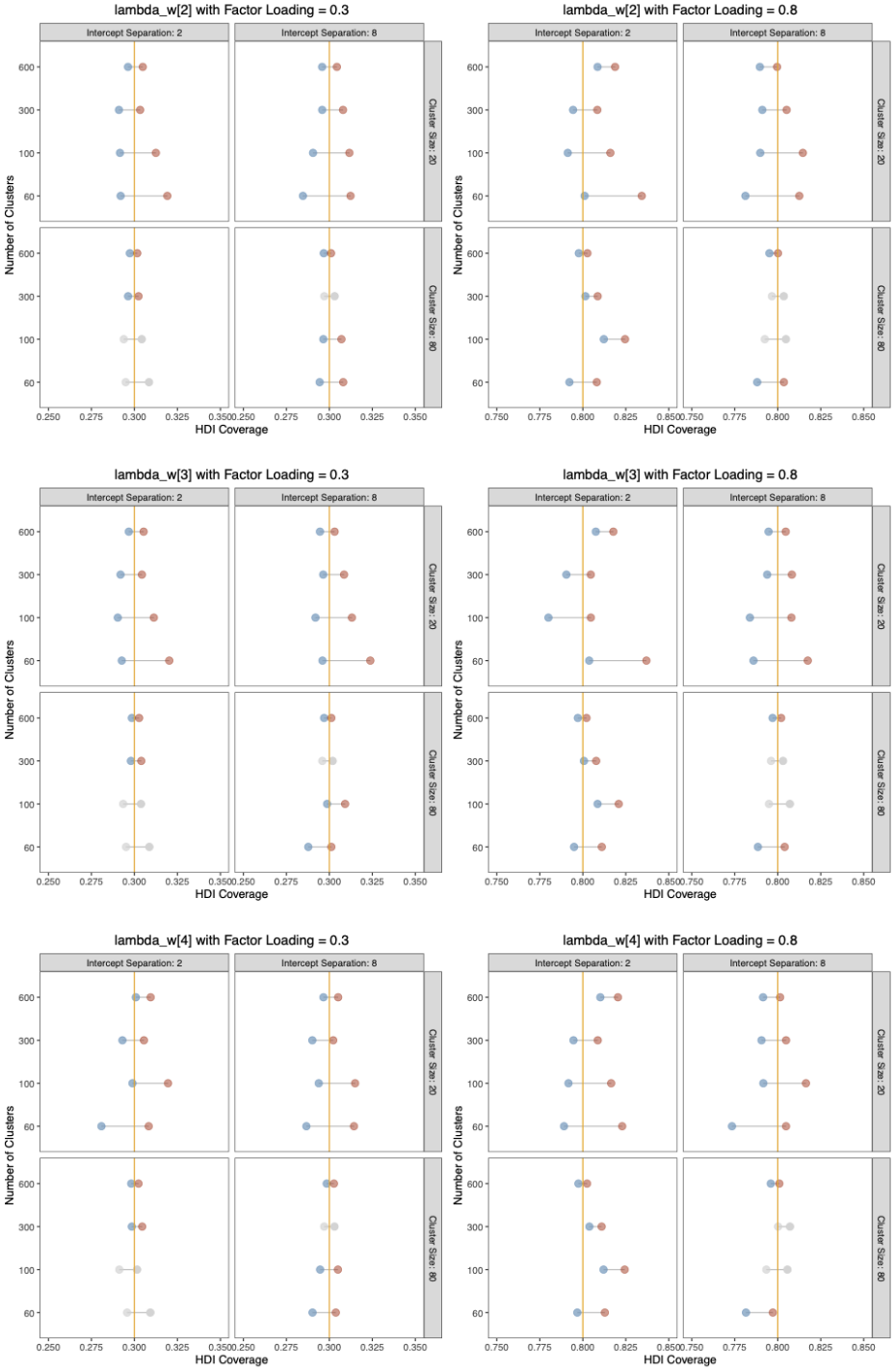


Figure B9

95% HPDI Coverage for Item Intercepts $\boldsymbol{\alpha}$ *in Class 1*

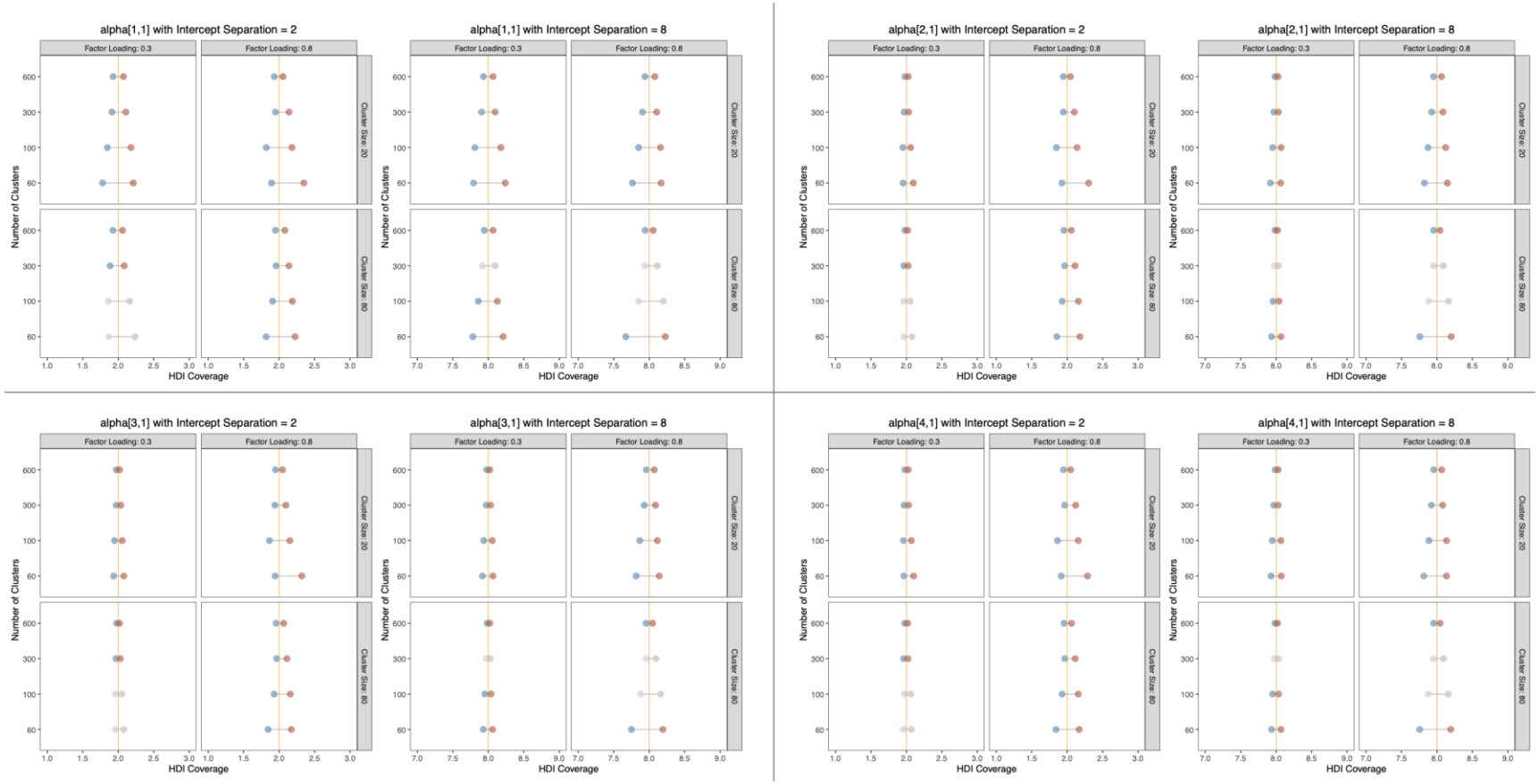


Figure B10

95% HPDI Coverage for Item Intercepts $\boldsymbol{\alpha}$ *in Class 2*

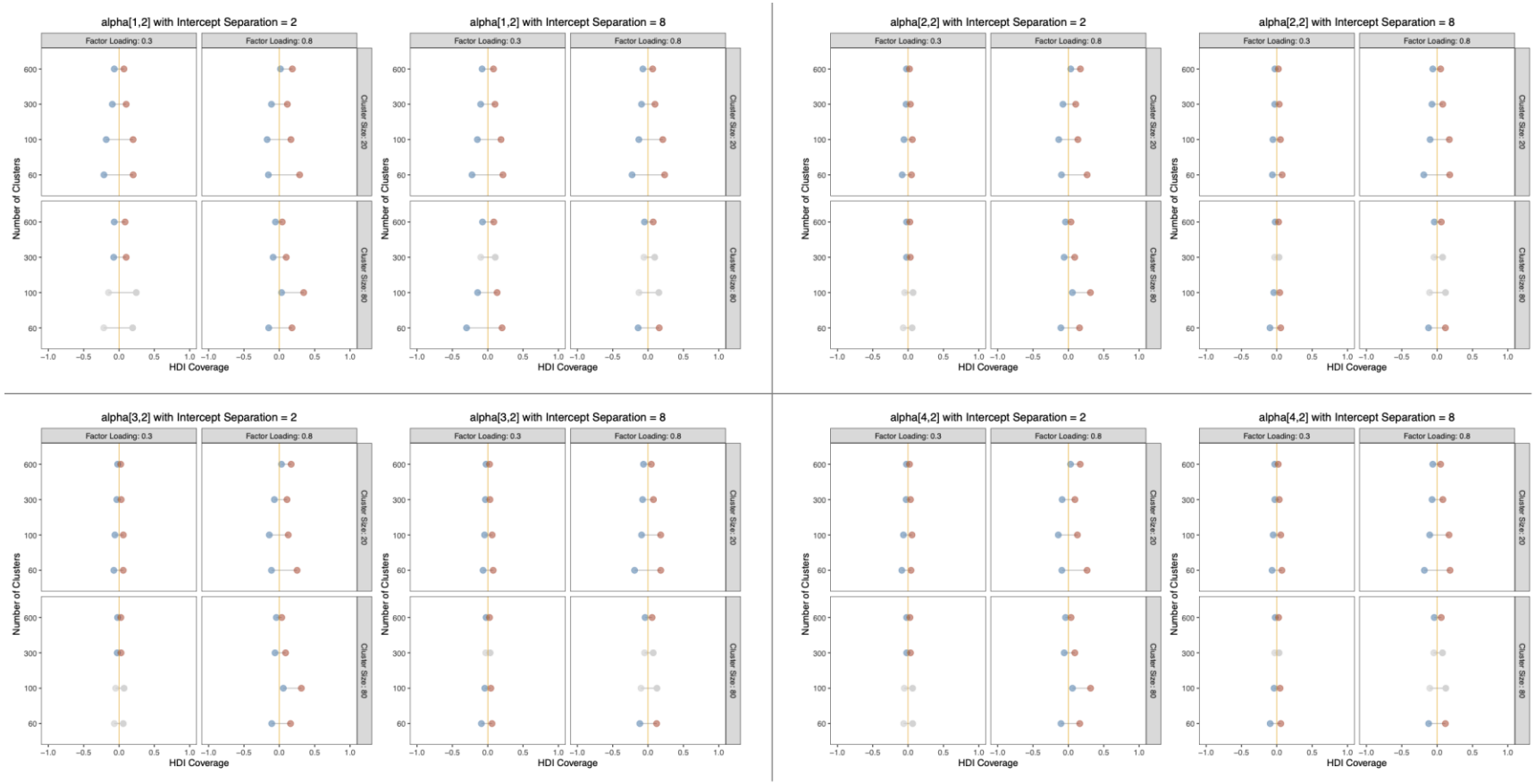


Figure B11

Relative Bias for Estimating the Probability of Belonging to Class 1 with Naïve Mean Score MMSEM-RV

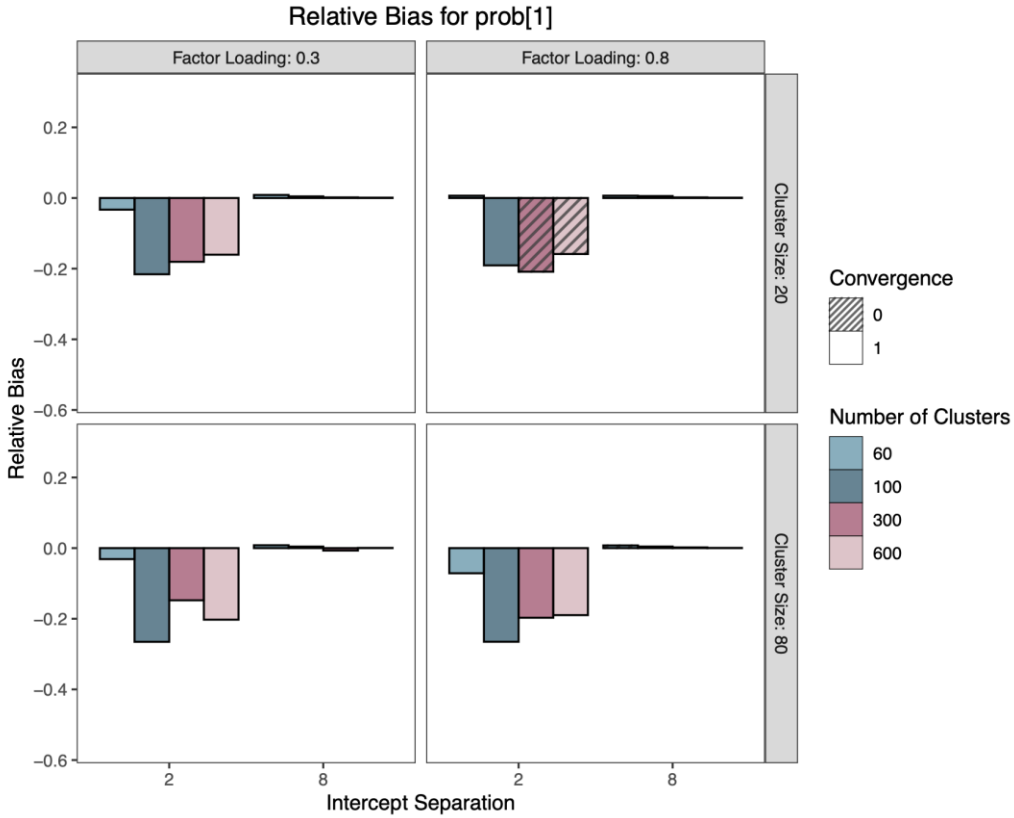
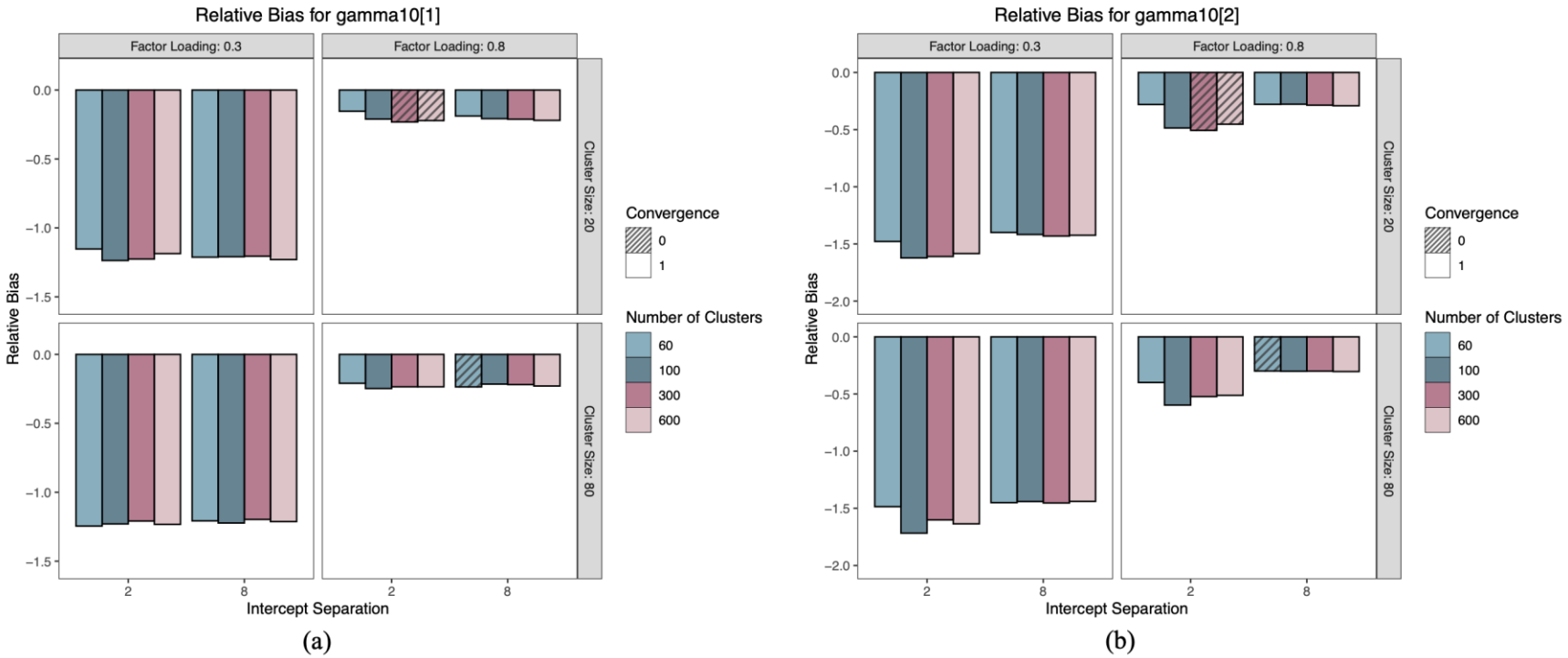


Figure B12

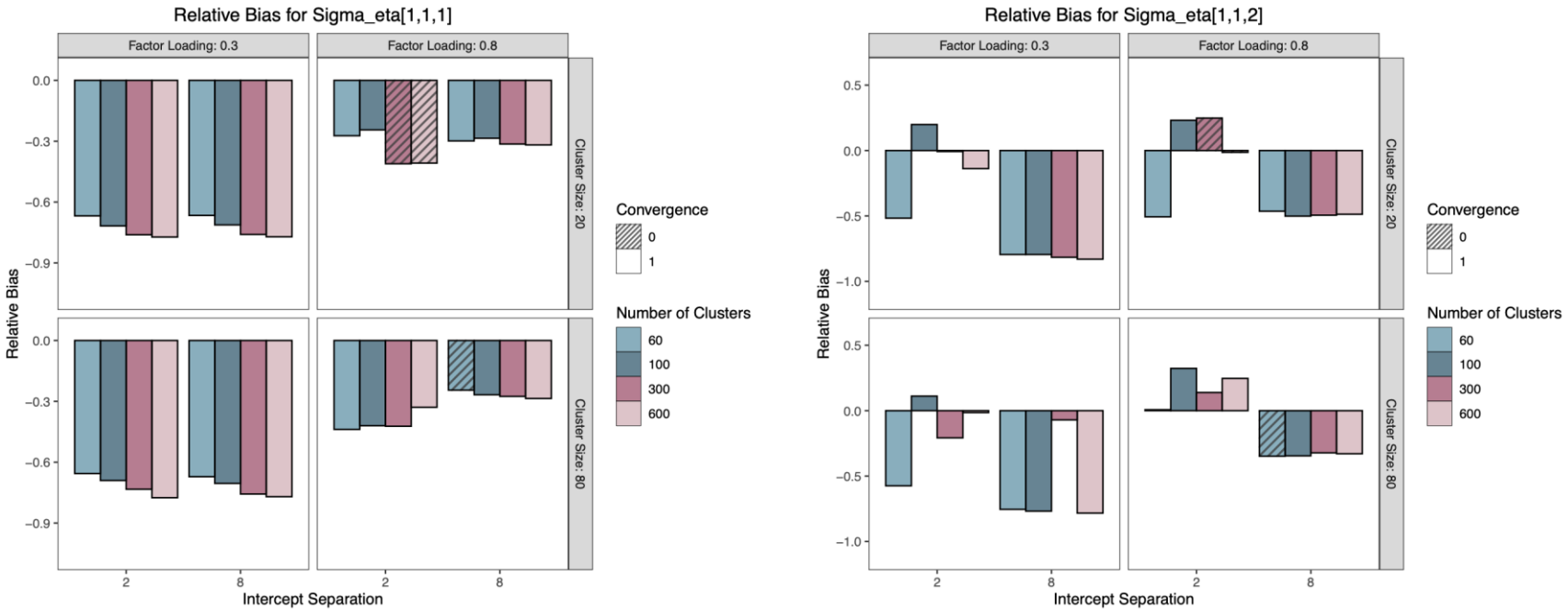
Relative Bias for Estimating Average Within-Level Log Variance in Class 1 and Class 2 with Naïve Mean Score MMSEM-RV



Note. (a) Relative bias for estimating γ_{101}. (b) Relative bias for estimating γ_{102}.

Figure B13

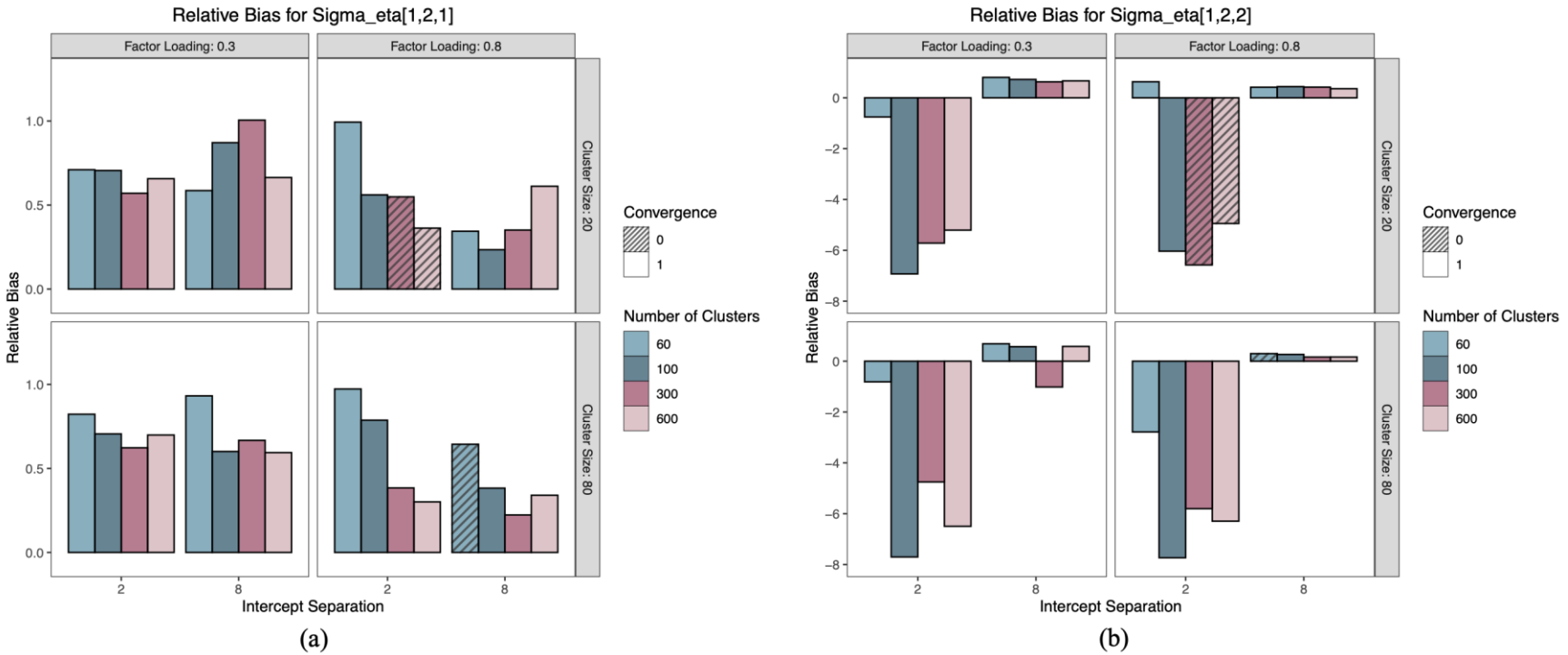
Relative Bias for Estimating the Between-Level Latent Construct Variance in Class 1 and Class 2 with Naïve Mean Score MMSEM-RV



Note. (a) Relative bias for estimating ϕ_{01}. (b) Relative bias for estimating ϕ_{02}.

Figure B14

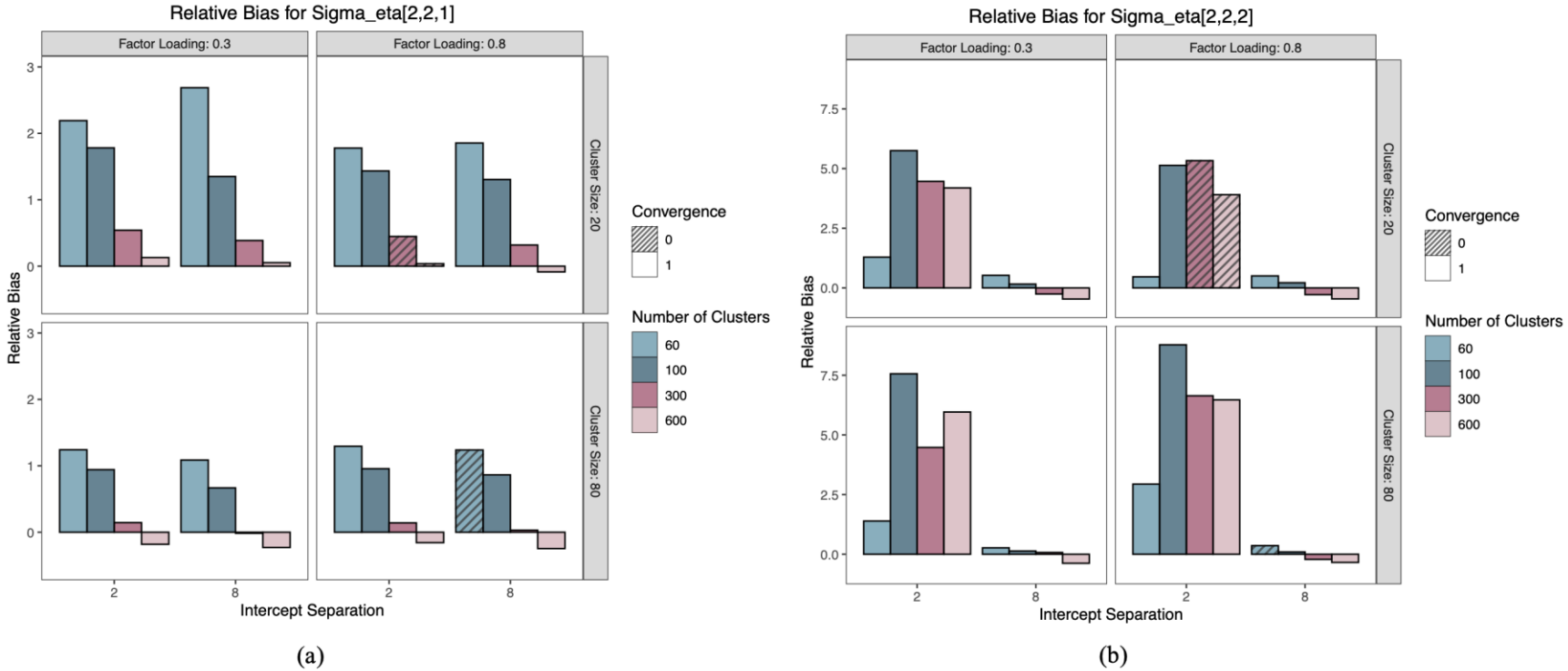
Relative Bias for Estimating the Between-Level Random Effects Covariance in Class 1 and Class 2 with Naïve Mean Score MMSEM-RV



Note. (a) Relative bias for estimating ϕ_{101}. (b) Relative bias for estimating ϕ_{102}.

Figure B15

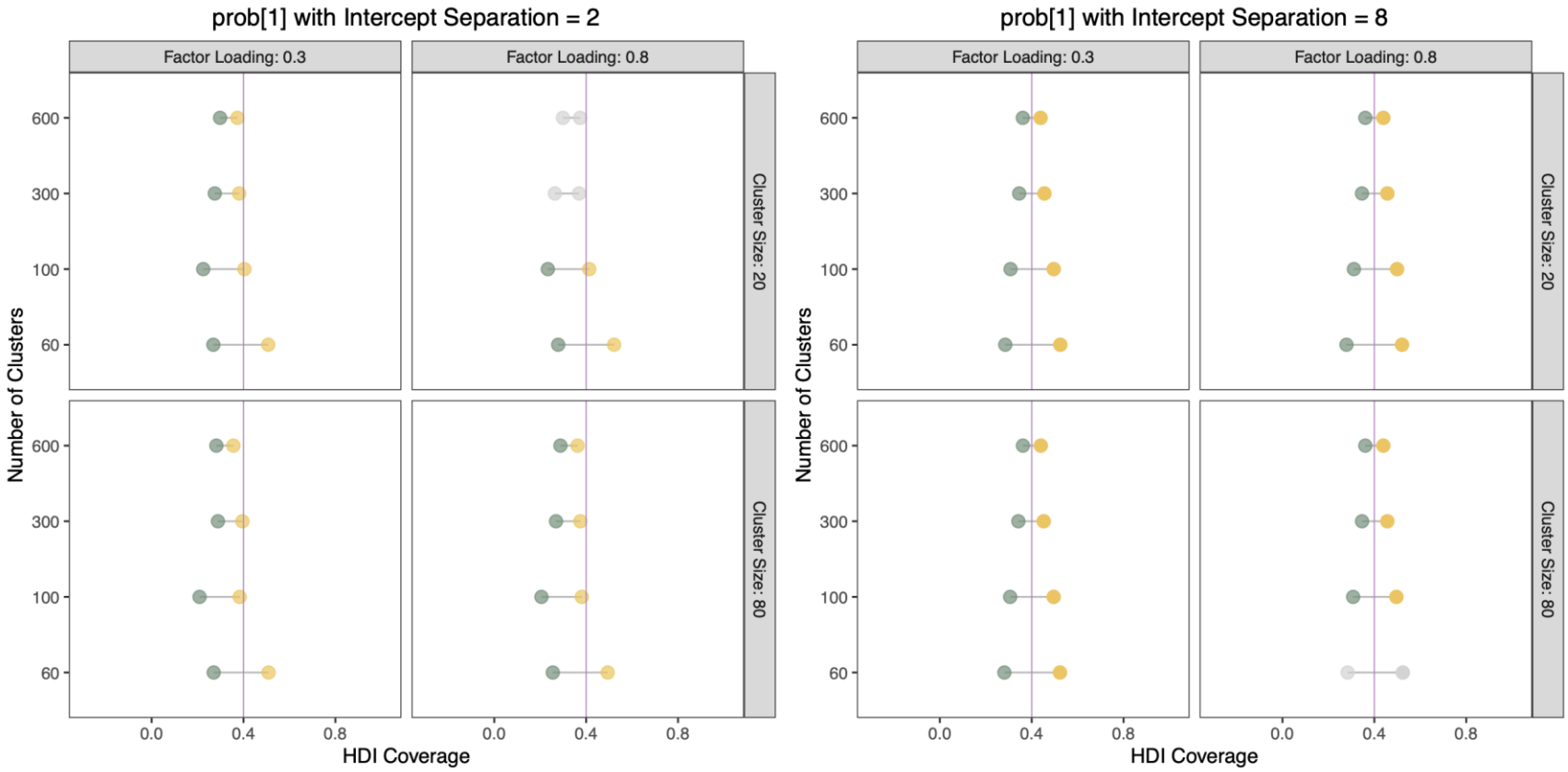
Relative Bias for Estimating Variance of the Within-Level Log Variance in Class 1 and Class 2 with Naïve Mean Score MMSEM-RV



Note. (a) Relative bias for estimating ϕ_{11}. (b) Relative bias for estimating ϕ_{12}.

Figure B16

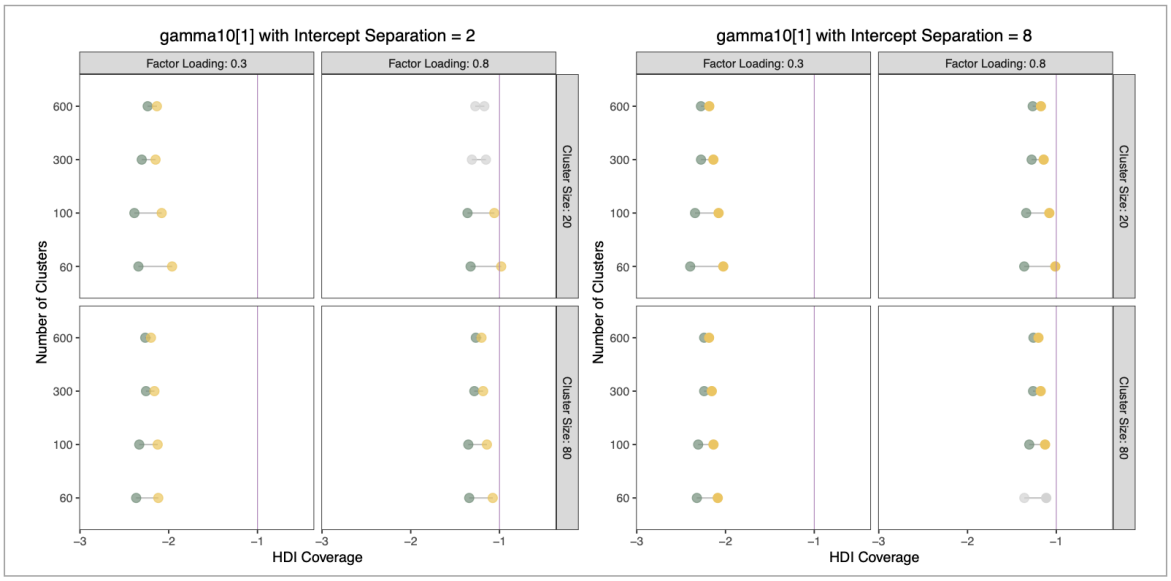
95% HPDI Coverage for Estimating the Probability of Belonging to Class 1 with Naïve Mean Score MMSEM-RV



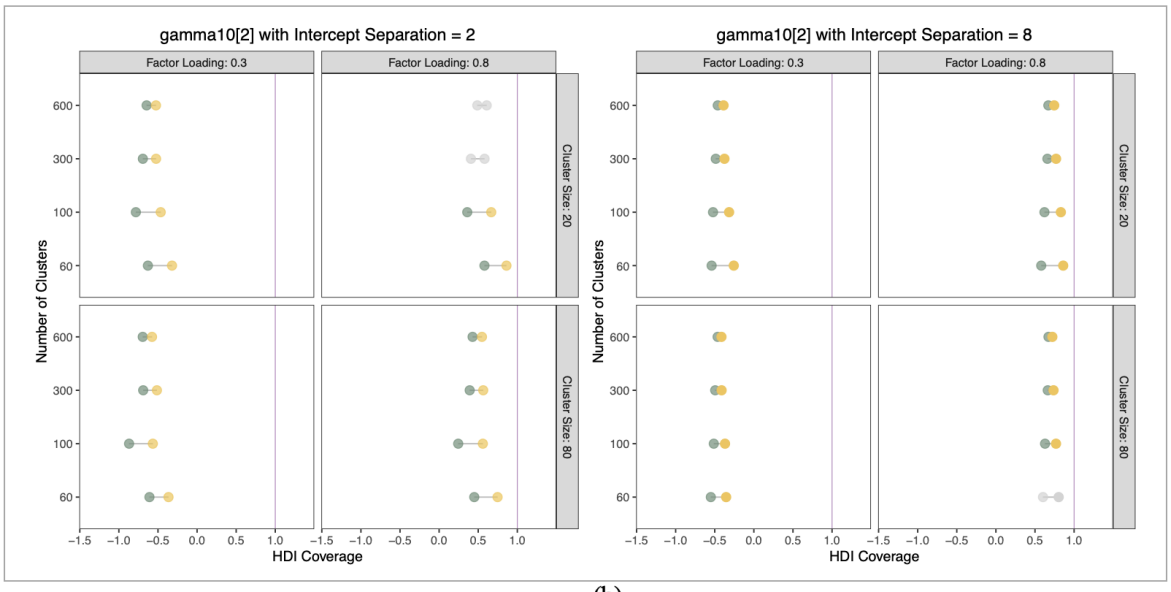
Note. Conditions that did not reach convergence are marked in light grey. Population value of the parameter is indicated by the purple line.

Figure B17

95% HPDI Coverage for Estimating Average Within-Level Log Variance in Class 1 and Class 2 with Naïve Mean Score MMSEM-RV



(a)

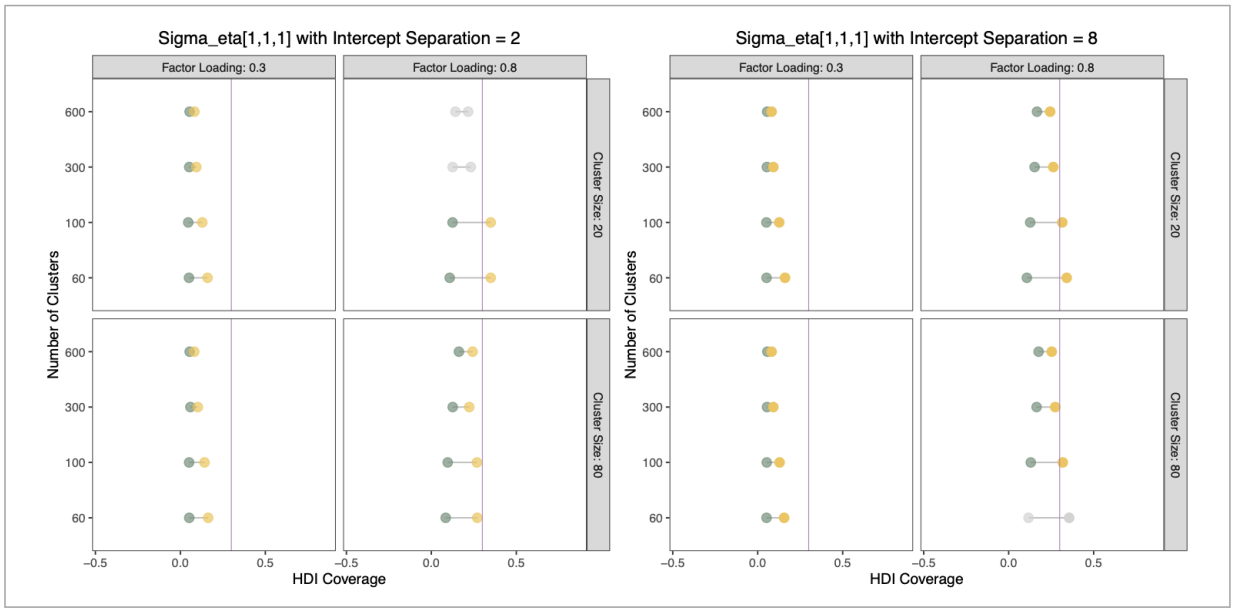


(b)

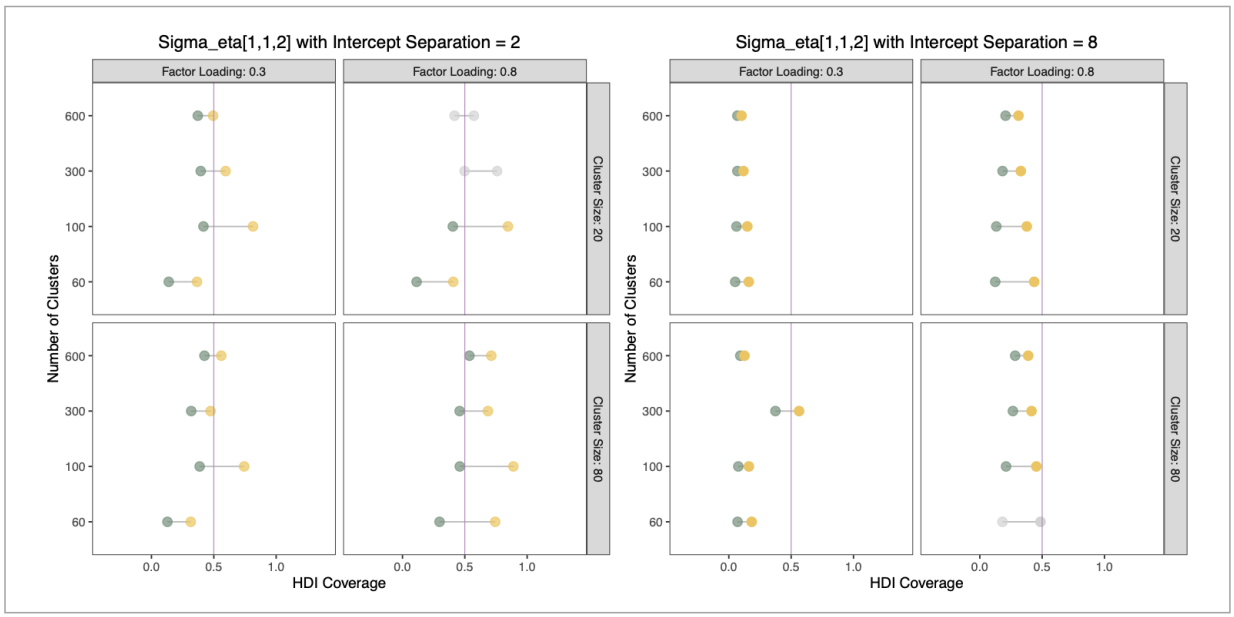
Note. (a) HPDI for estimating γ_{101}. (b) HPDI for estimating γ_{102}. Conditions that did not reach convergence are marked in light grey. Population value of the parameter is indicated by the purple line.

Figure B18

95% HPDI Coverage for Estimating the Between-Level Latent Construct Variance in Class 1 and Class 2 with Naïve Mean Score MMSEM-RV



(a)

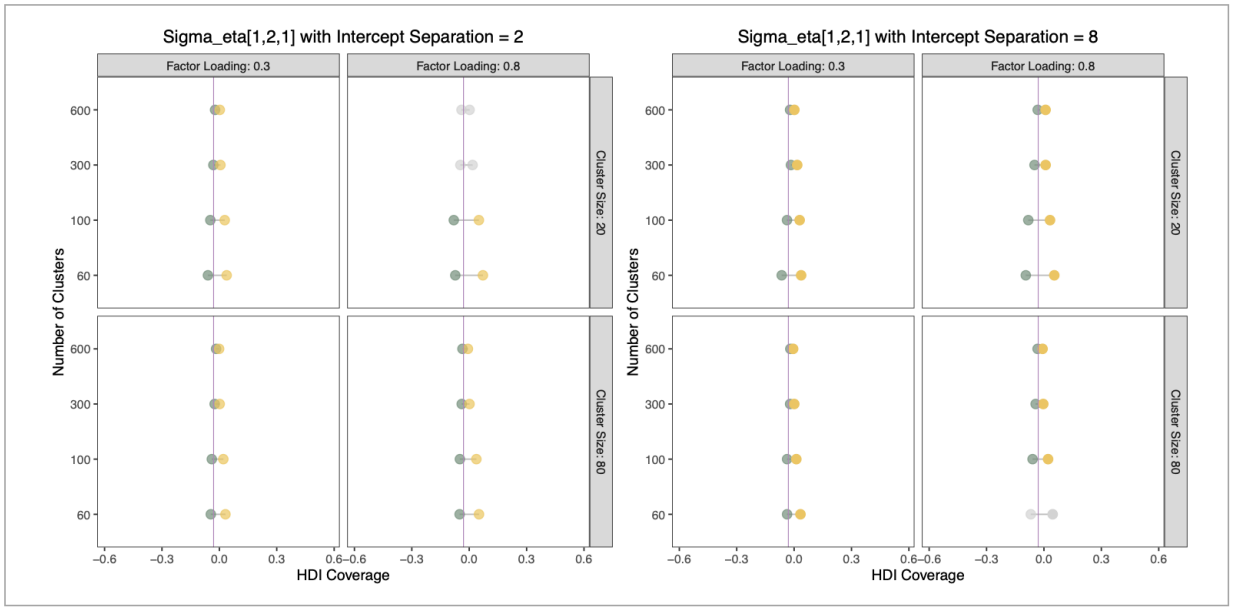


(b)

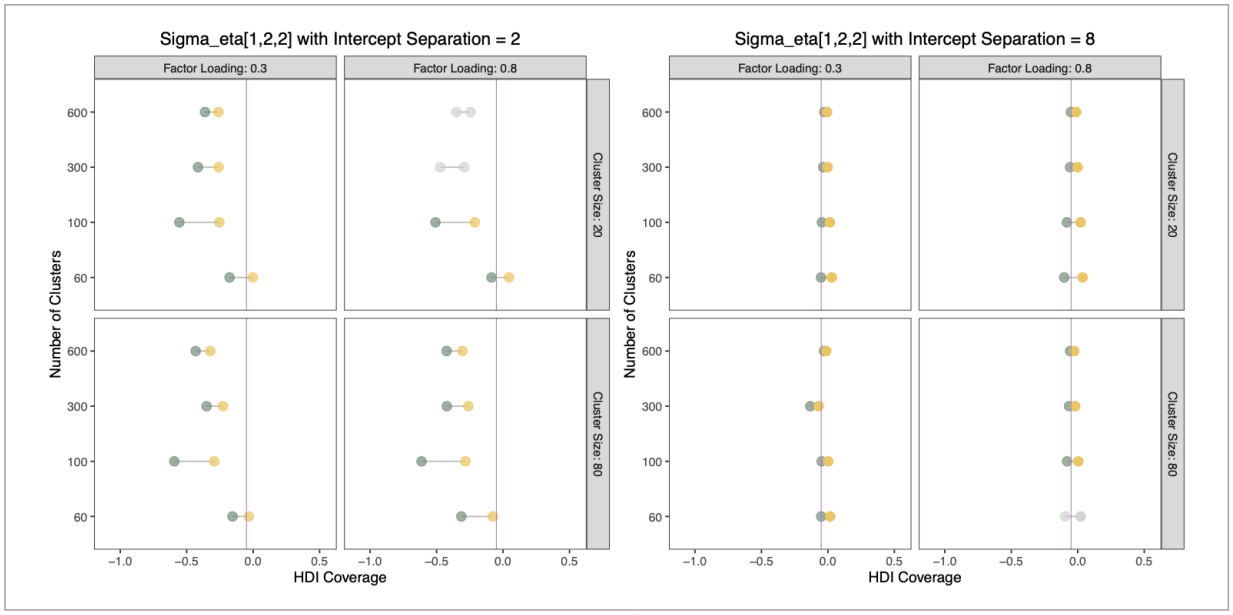
Note. (a) HPDI for estimating ϕ_{01}. (b) HPDI for estimating ϕ_{02}. Conditions that did not reach convergence are marked in light grey. Population value of the parameter is indicated by the purple line.

Figure B19

95% HPDI Coverage for Estimating the Between-Level Random Effects Covariance in Class 1 and Class 2 with Naïve Mean Score MMSEM-RV



(a)

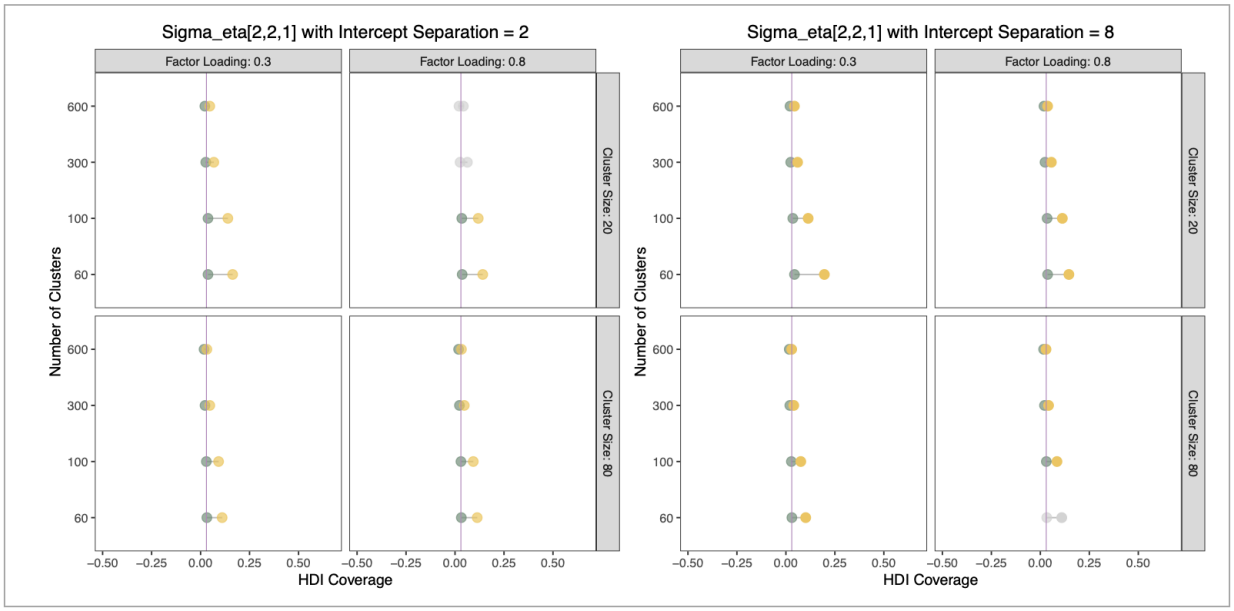


(b)

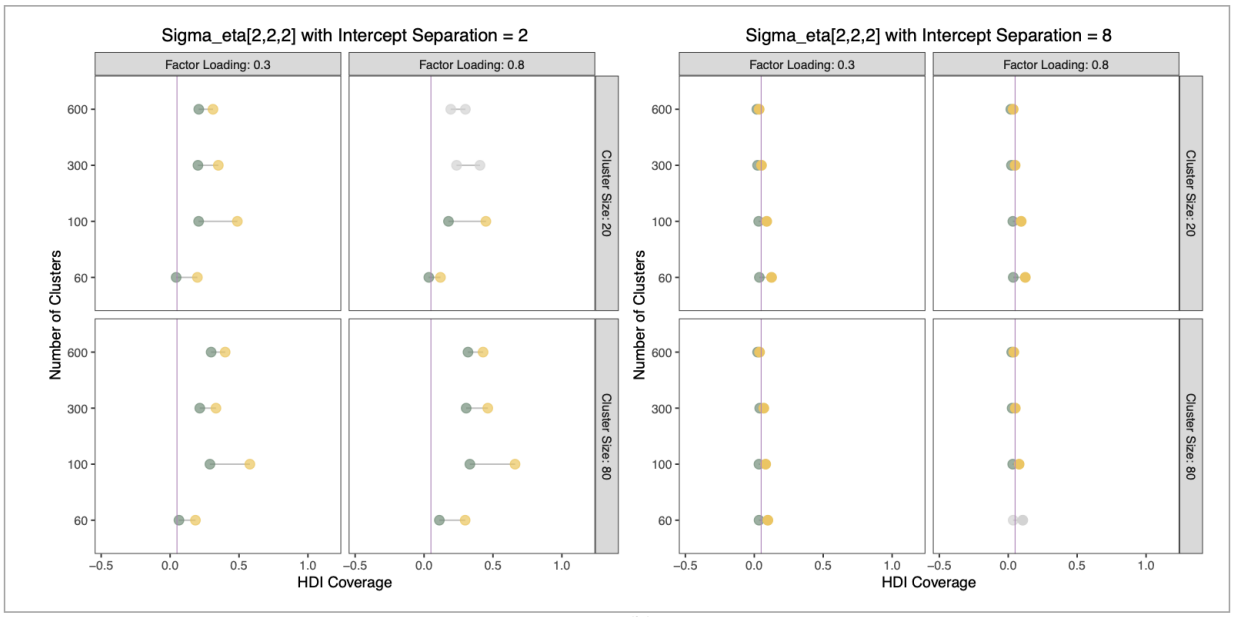
Note. (a) HPDI for estimating ϕ_{101}. (b) HPDI for estimating ϕ_{102}. Conditions that did not reach convergence are marked in light grey. Population value of the parameter is indicated by the purple line.

Figure B20

95% HPDI Coverage for Estimating Variance of the Within-Level Log Variance in Class 1 and Class 2 with Naïve Mean Score MMSEM-RV



(a)



(b)

Note. (a) HPDI for estimating ϕ_{11}. (b) HPDI for estimating ϕ_{12}. Conditions that did not reach convergence are marked in light grey. Population value of the parameter is indicated by the purple line.

Figure B21

Running Mean Plots for Parameter Estimates with Latent Construct MMSEM-RV across Simulation Replications under Condition $J = 60$, $n_j = 20$, $\alpha_1 = 2$, $\lambda = 0.3$

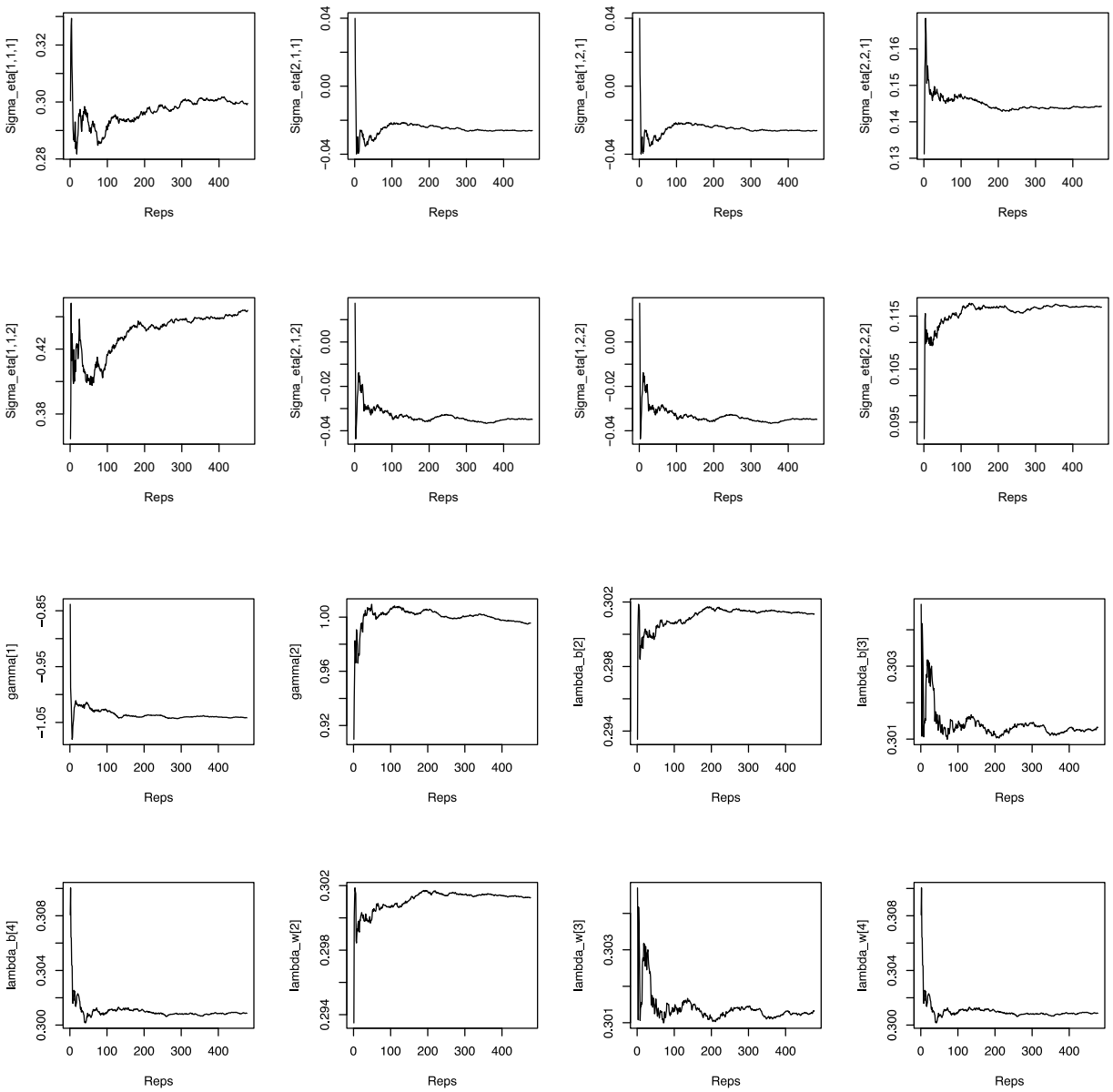


Figure B21 Continued

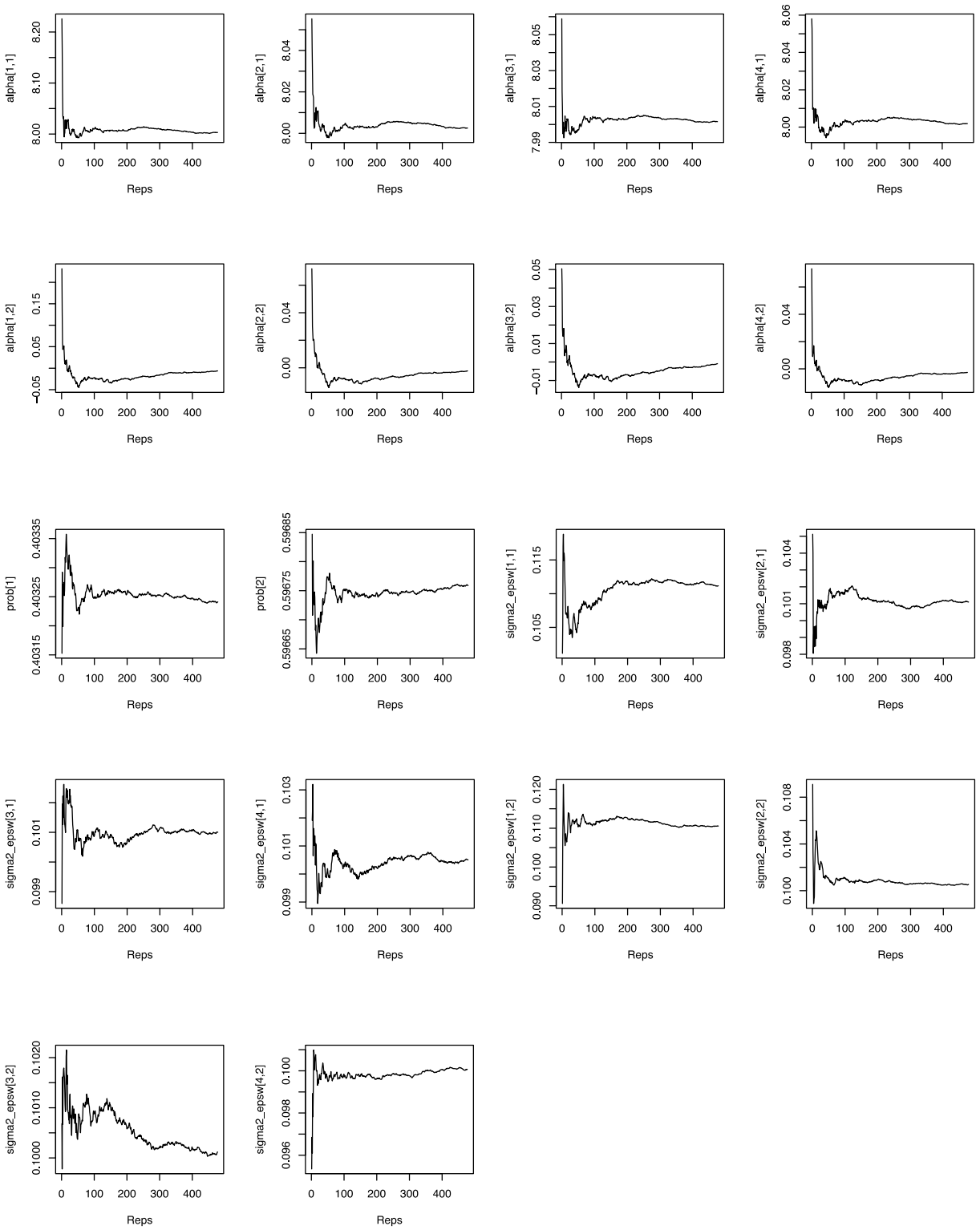


Figure B22

Running Mean Plots for Parameter Estimates with Naïve Mean Score MMSEM-RV across Simulation Replications under Condition $J = 60$, $n_j = 20$, $\alpha_1 = 2$, $\lambda = 0.3$

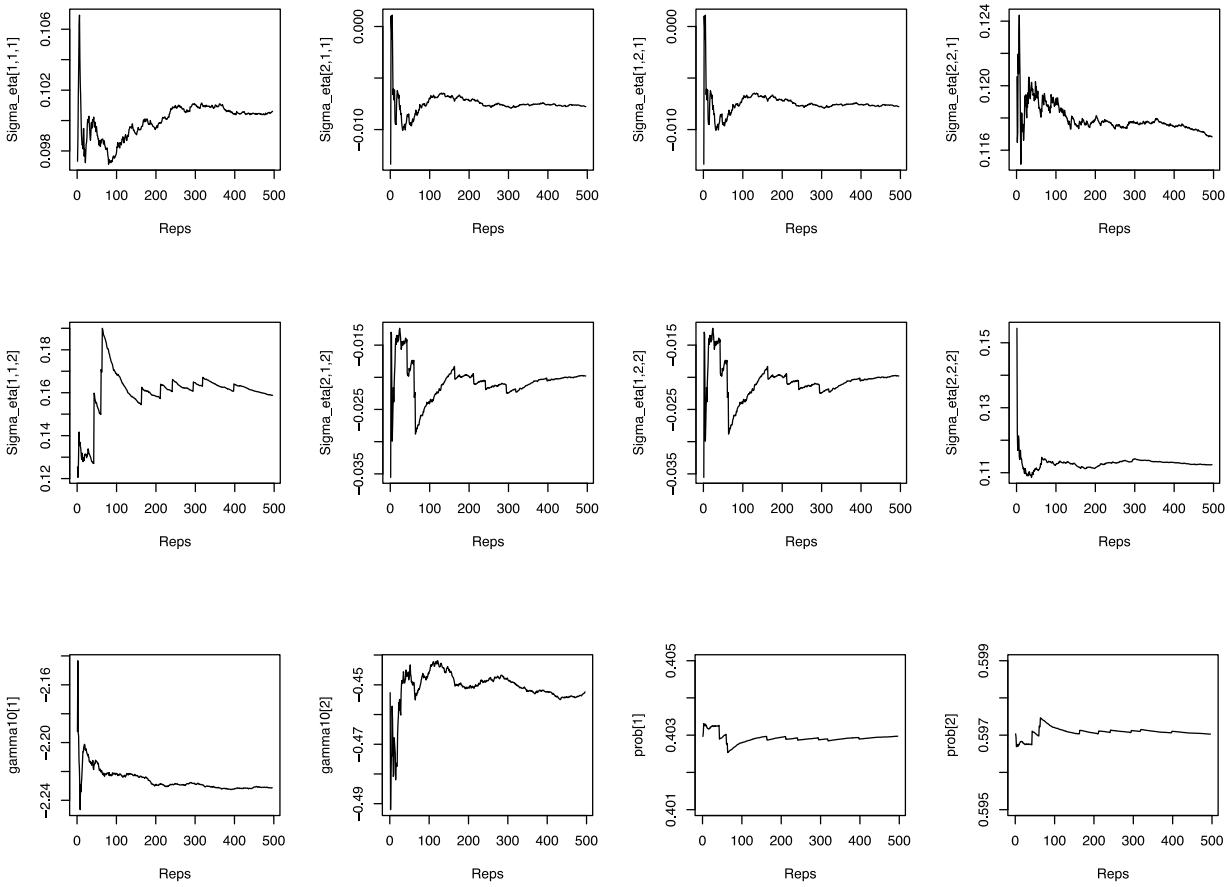


Figure B23

Running Mean Plots for Parameter Estimates with Latent Construct MMSEM-RV across Simulation Replications under Condition $J = 600$, $n_j = 80$, $\alpha_1 = 8$, $\lambda = 0.8$

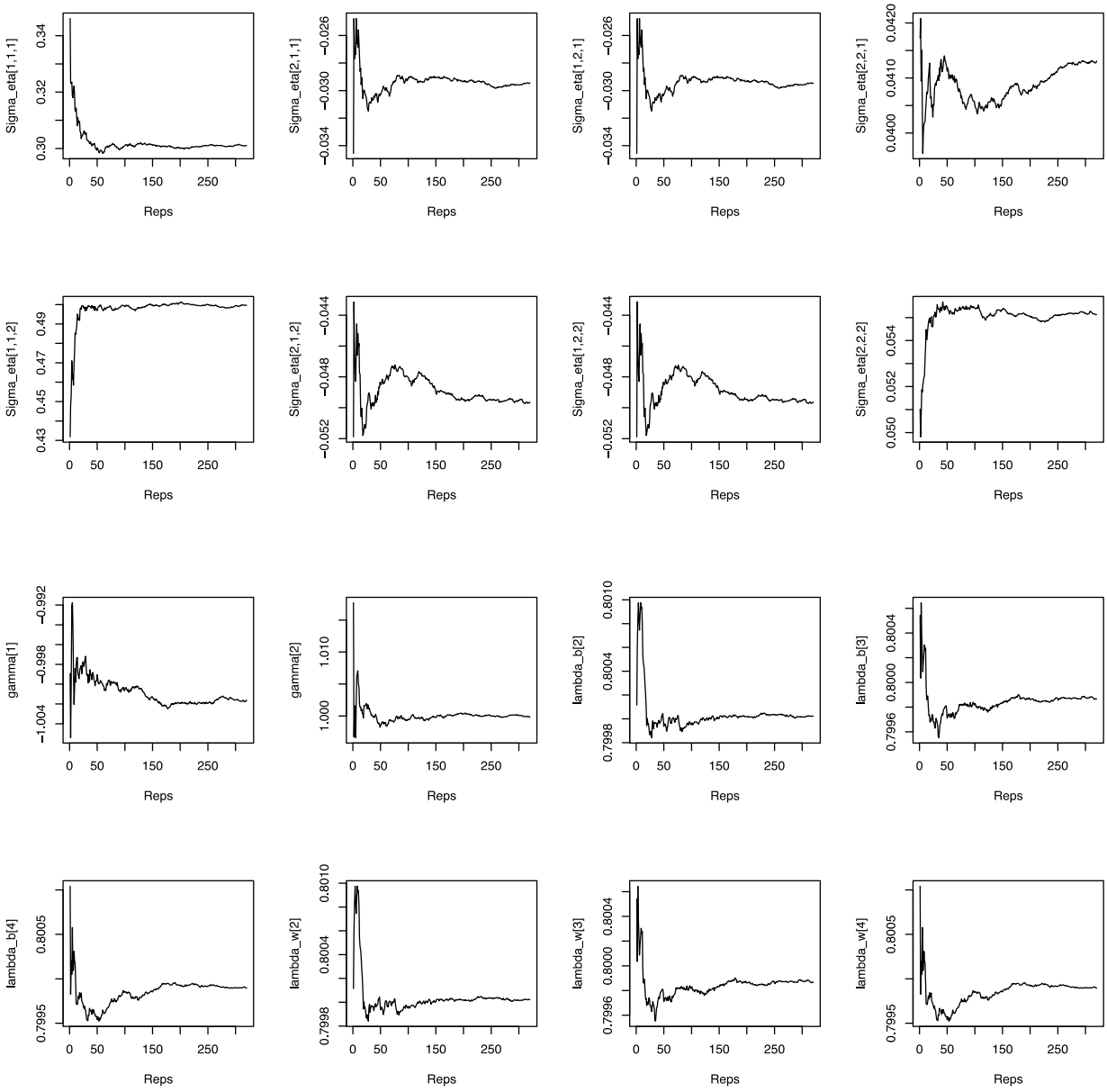


Figure B23 Continued

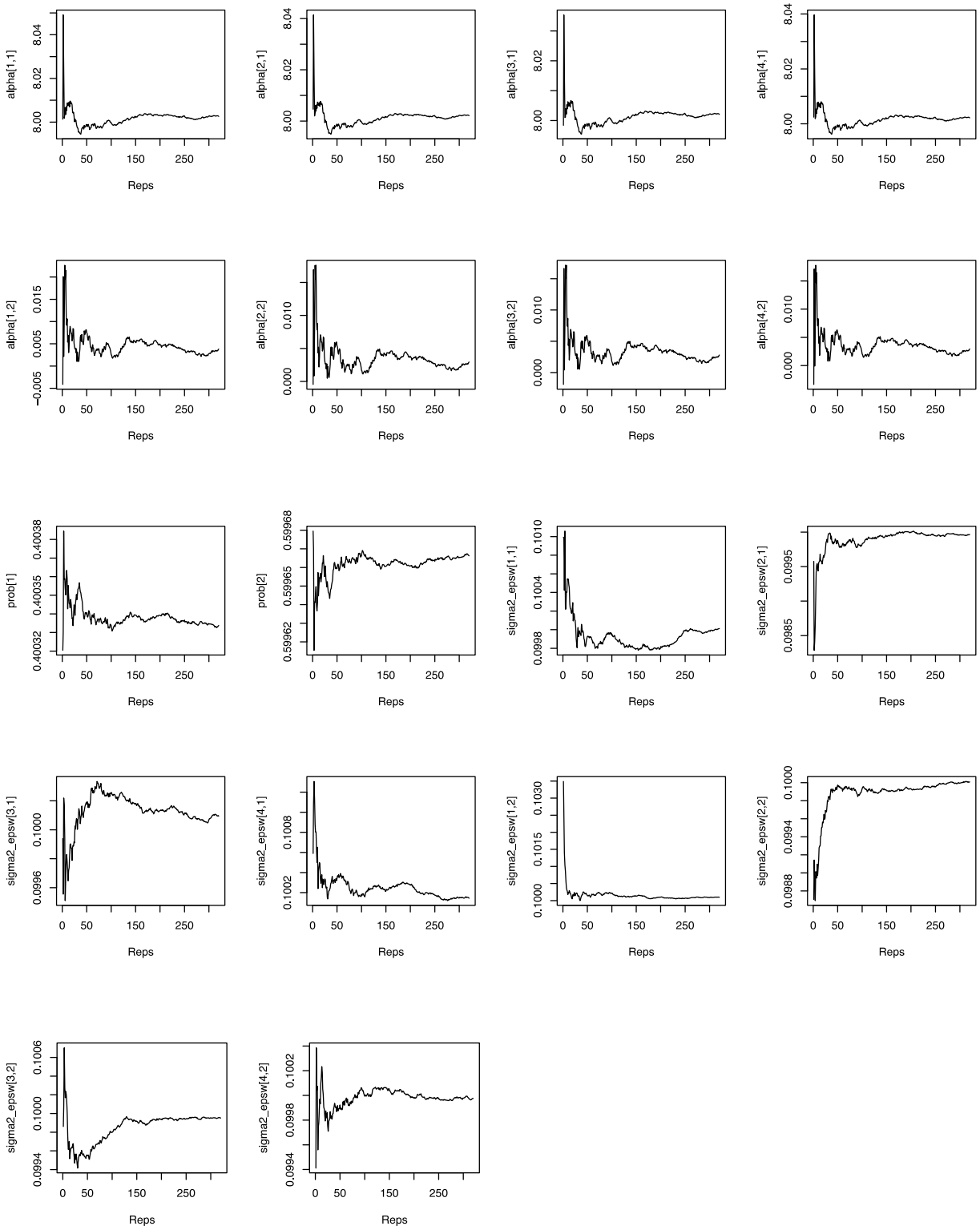


Figure B24

Running Mean Plots for Parameter Estimates with Naïve Mean Score MMSEM-RV across Simulation Replications under Condition $J = 600$, $n_j = 80$, $\alpha_1 = 8$, $\lambda = 0.8$

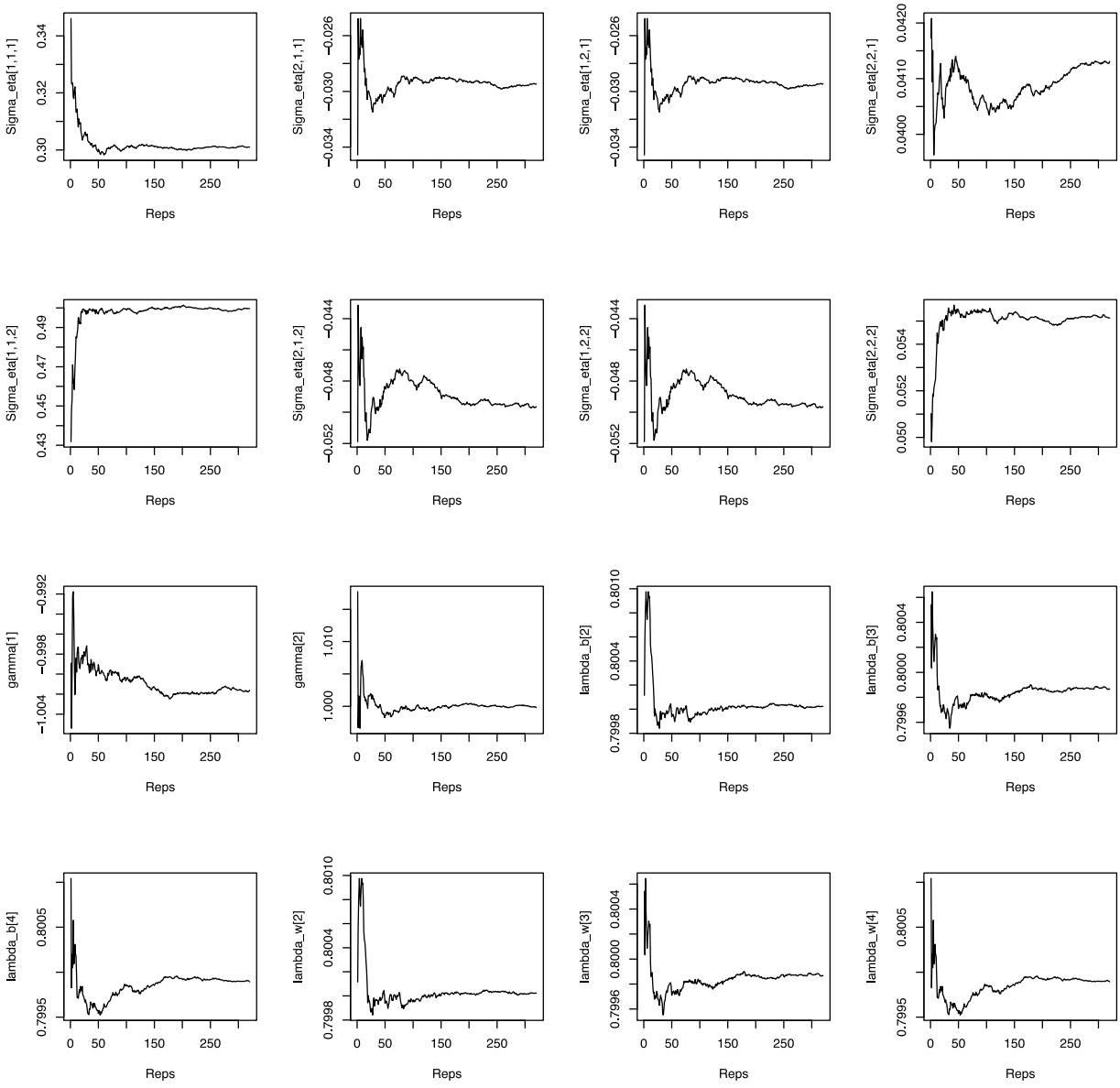


Figure B24 Continued

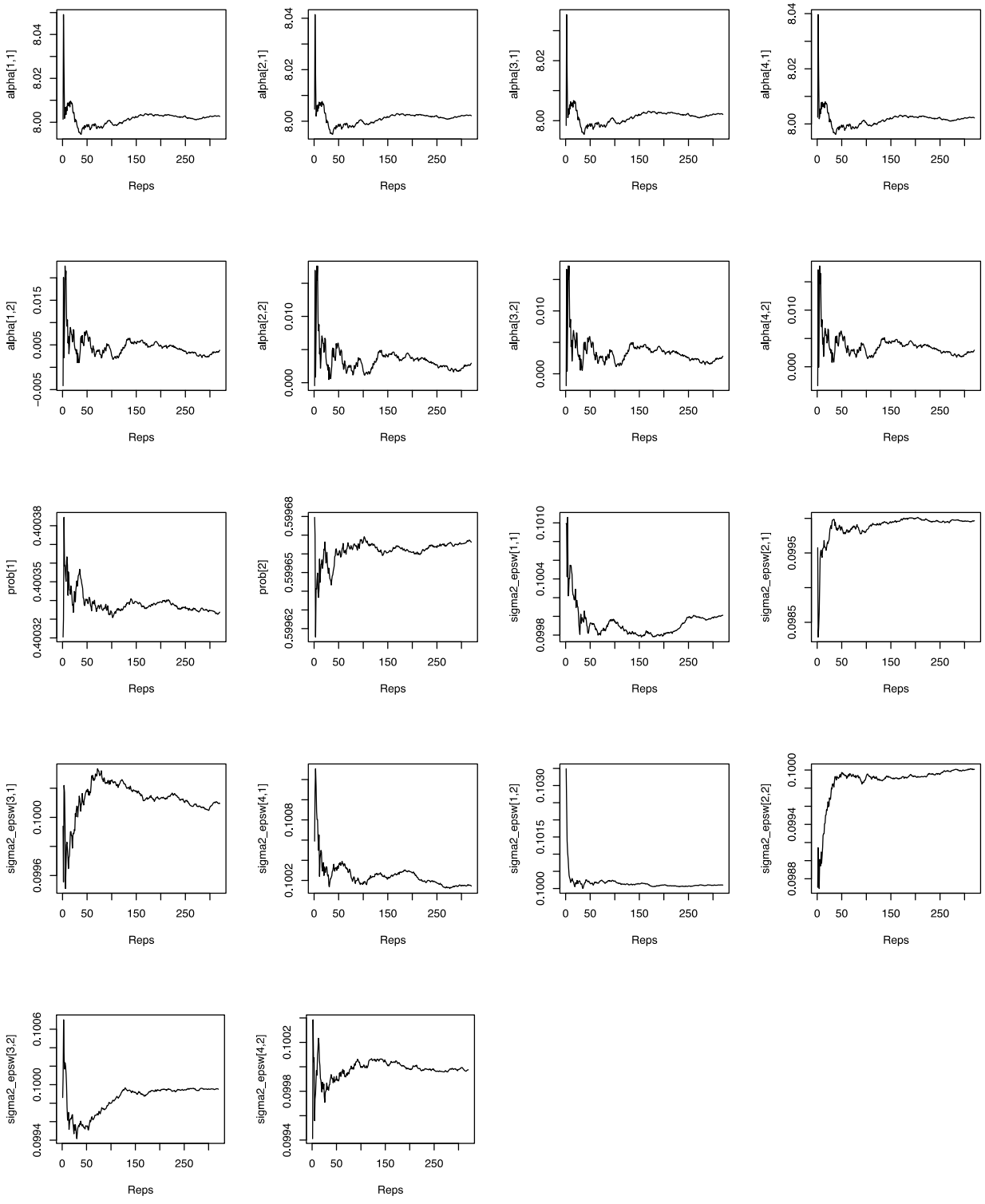


Figure B25

Point Estimates with Standard Error for Peripheral Parameters with Latent Construct MMSEM-RV under Condition A

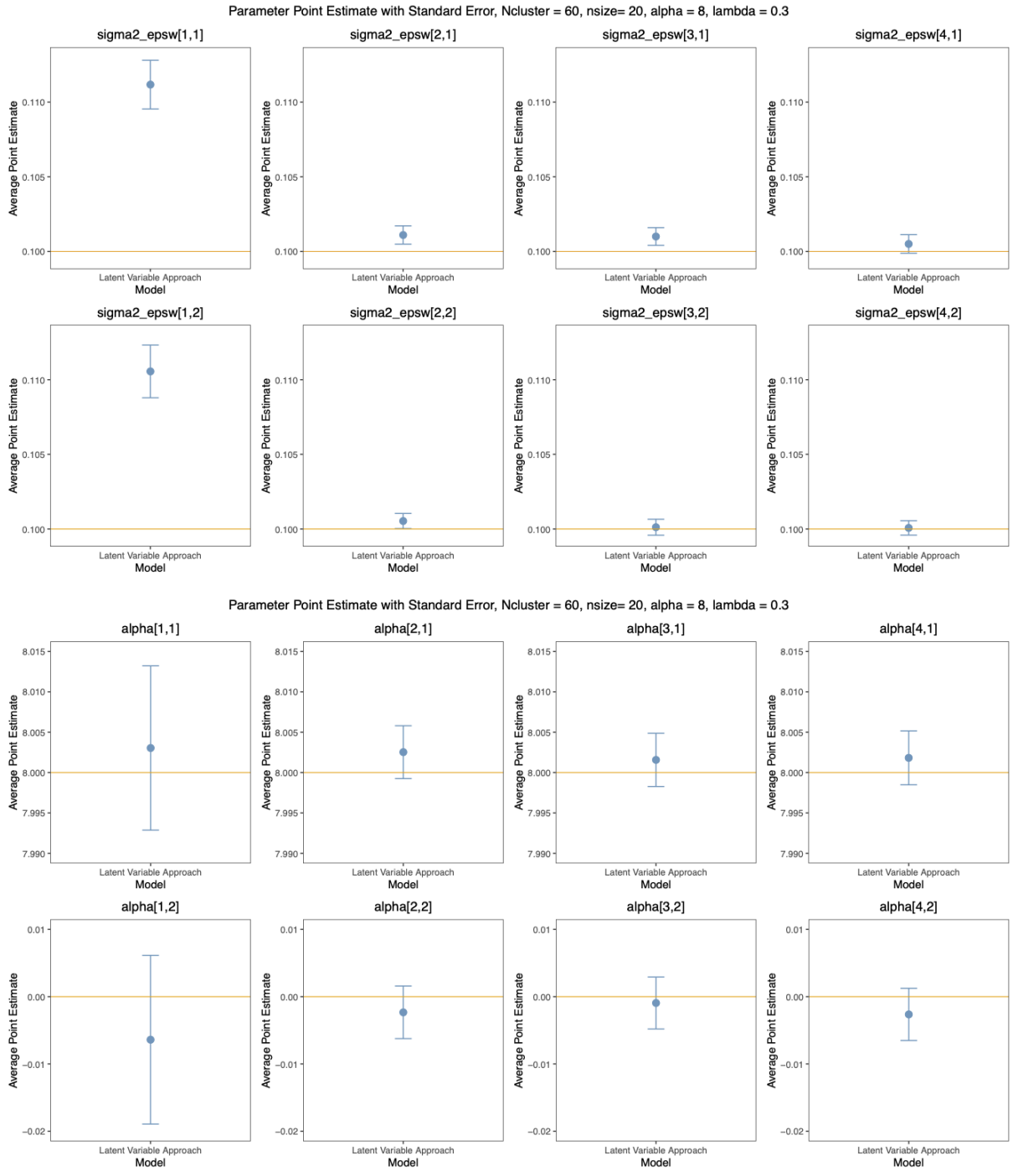


Figure B25 Continued

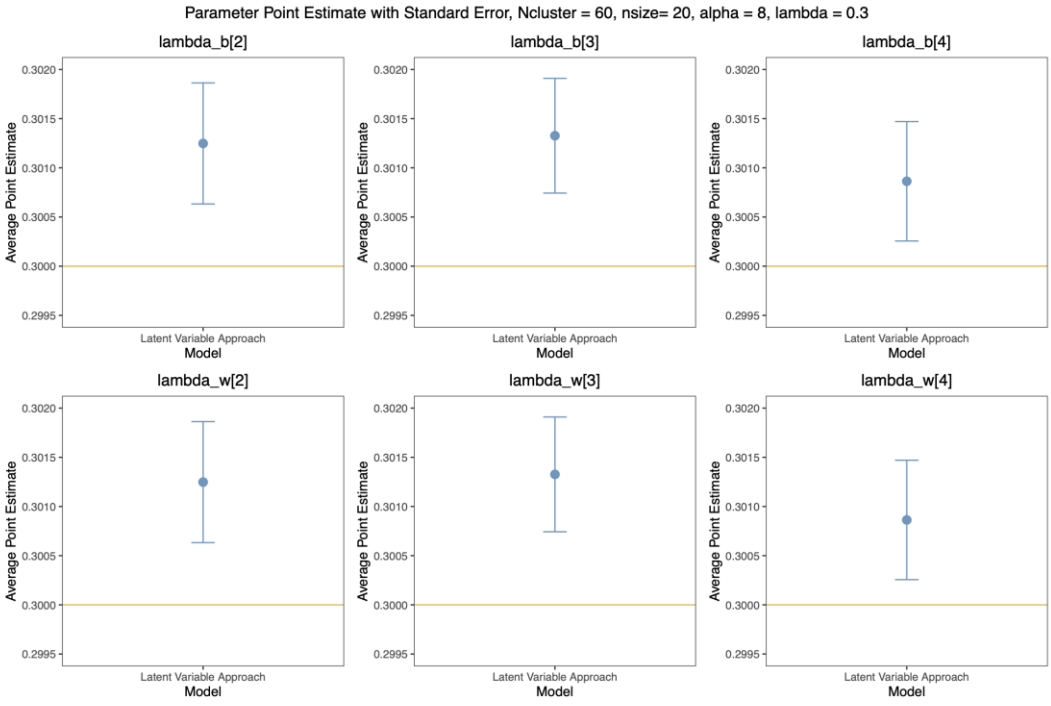


Figure B26

Point Estimates with Standard Error for Peripheral Parameters with Latent Construct MMSEM-RV under Condition B

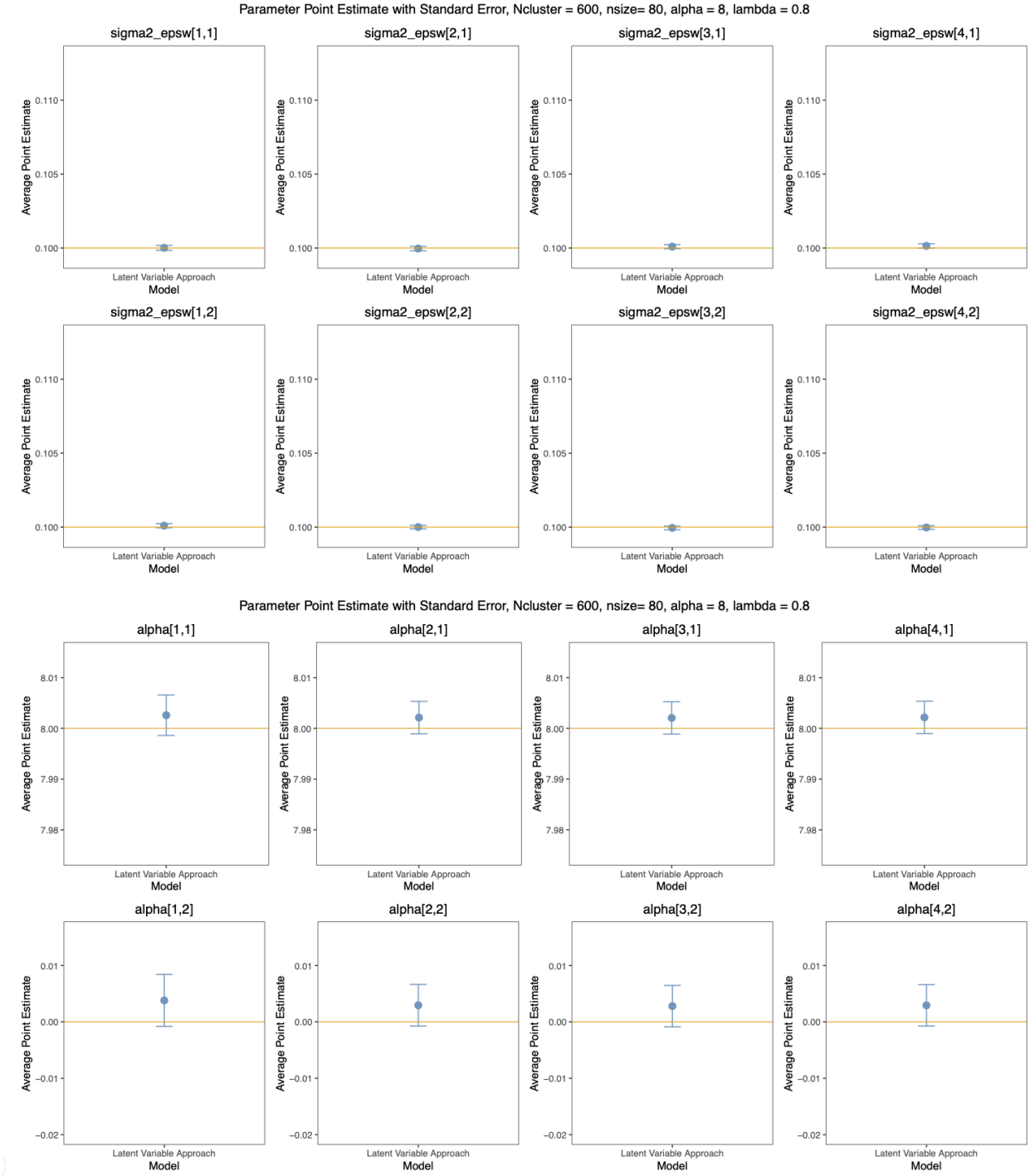
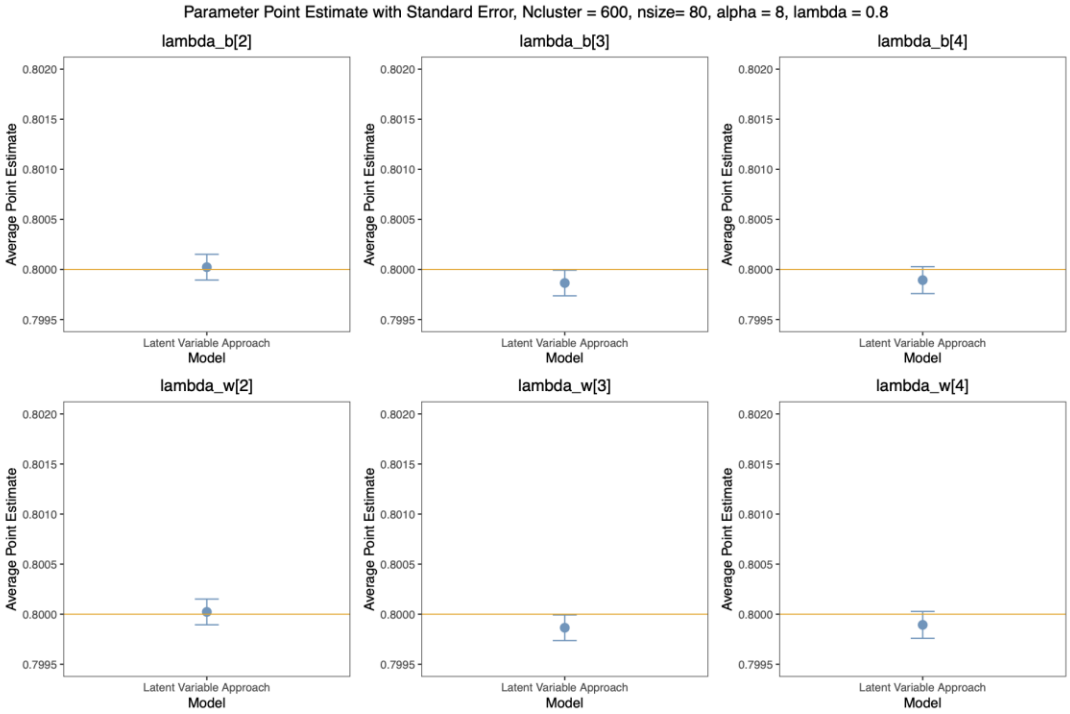


Figure B26 Continued



Appendix C. BUGS Code for Fitting the Latent Construct MMSEM-RV

```
# Example BUGS code used for fitting a 2-class latent construct MMSEM-
RV

model{

  #-------------- level 1 model --------------


  for (i in 1:nobs){

# for ith observation

  z.ij[i] ~ dnorm(0, sigma2_inv[ cluster[i] ])

# within-level latent construct, precision is specific within each
# cluster

    for (p in 1:nitems){

# for pth observed item


      imean[i, p] <- alpha[p, class[ cluster[i]] ] +
lambda_b[p]*z.j[ cluster[i] ] + epsb[p, class[ cluster[i] ] ] +
lambda_w[p]*z.ij[i]

       y[i, p] ~ dnorm(imean[i, p], sigma2inv_epsw[ p,
class[ cluster[i]] ])
    }
  }


  #-------------- level 2 model --------------

  for (j in 1:nclusters){

    class[j] ~ dcat(prob[])
    # latent class membership

    z.j[j] <- eta[j, 1]
    # between-level latent construct

    logsigma2[j] <- gamma[ class[j] ] + eta[j, 2]
    # log variance of cluster j;

    # between-level random effects
    eta[j, 1:2] ~ dmnorm(rep(0, 2), Sigma_prec_eta[1:2, 1:2,
class[j] ])
```

```
  sigma2[j] <- exp(logsigma2[j])
  # within-level variance for each cluster

  sigma[j] <- pow(sigma2[j], 1/2)
  # within-level standard deviation for each cluster

  sigma2_inv[j] <- pow(sigma2[j], -1)
  # within-level precision parameter


}


#---------------- priors -----------------

prob[1:nclasses] ~ ddirch(alpha_gamma[])
# mixing proportions for a 2-class solution

 alpha_gamma[1] <- 1
 alpha_gamma[2] <- 1

# order constraints

gamma[1] ~ dnorm(-1, 1e-05)          # fixed effects of log variance
gamma[2] ~ dnorm(1, 1e-05) T(gamma[1], )

# loadings (not allowed to vary across latent classes)

lambda_w[1] <- 1                     # first loading
lambda_b[1] <- 1

for (p in 2:nitems){
lambda_w[p] ~ dnorm(0, 1e-07)        # free loadings
lambda_b[p] <- lambda_w[p]
# constrained to be equal across levels
}

# between-level random effects covariance
Sigma_prec_eta[1:2, 1:2, 1] ~ dwish(Imat1, 6)
Sigma_prec_eta[1:2, 1:2, 2] ~ dwish(Imat2, 6)


for (p in 1:nitems){

alpha[p, 1] ~ dnorm(8, 1e-01)
# informative prior for grand intercepts in latent class 1
alpha[p, 2] ~ dnorm(0, 1e-01)
# informative prior for grand intercepts in latent class 2

}
```

```
  for (g in 1:nclasses){

     for (p in 1:nitems){

     sigma2inv_epsw[p, g] ~ dgamma(3, 0.5)
     sigma2_epsw[p, g] <- 1/sigma2inv_epsw[p, g]
     # level-1 residual variances

     sigma2_epsb[p, g] <- 0
     epsb[p, g] <- 0
     # level-2 residual variances constrained to be zero


     }


  # level-2 variance components

  Sigma_eta[1:2, 1:2, g] <- inverse(Sigma_prec_eta[1:2, 1:2, g])


  #----------- inference about the variance
  # lognormal distribution

  # mean variance
  mean.v[g] <- exp(gamma[g] + Sigma_eta[2, 2, g]/2)

  # variance of variance
  var.v[g] <- (exp(Sigma_eta[2, 2, g]) - 1)*exp(2*gamma[g] +
Sigma_eta[2, 2, g])

  }

  var.ratio <- Sigma_eta[2, 2, 1]/Sigma_eta[2, 2, 2]
  mean.diff <- gamma[1] - gamma[2]

}
```

Appendix D. Estimation Results of the Naive Mean Score MMSEM-RV for the Applied Illustrative Example

Table D1

Results for Fitting a 2-Class Naïve Mean Score MMSEM-RV to the Illustrative Example

		K = 2	
		Posterior Mean	95% HPDI
Parameters	γ_{101}	-.629	(-.698, -.560)
	γ_{102}	253.428	(-.603, 624.494)
	π_1	.994	(.982, 1.000)
	ϕ_{01}	.051	(.036, .067)
	ϕ_{101}	-.043	(-.062, -.024)
	ϕ_{11}	.092	(.056, .131)
	ϕ_{02}	.330	(.046, .883)
	ϕ_{102}	-.000	(-.409, .424)
	ϕ_{12}	.338	(.049, .889)
	α_1	3.214	(3.171, 3.258)
	α_2	1.898	(-4.367, 8.116)

References

Adolph, K. E., Cole, W. G., & Vereijken, B. (2015). Intraindividual variability in the development of motor skills in childhood. In M. Diehl, K. Hooker, & M. Sliwinski (Eds.), *Handbook of intraindividual variability across the life span* (pp. 59–83). Routledge/Taylor & Francis Group.

Aitkin, M. (1987). Modelling variance heterogeneity in normal regression Using GLIM. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *36*(3), 332–339. https://doi.org/10.2307/2347792

Asparouhov, T., Hamaker, E. L., & Muthén, B. (2018). Dynamic structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(3), 359–388. https://doi.org/10.1080/10705511.2017.1406803

Asparouhov, T., & Muthén, B. (2007). Computationally efficient estimation of multilevel high-dimensional latent variable models. *2007 JSM Meeting in Salt Lake City, Utah, Section on Statistics in Epidemiology*, 2531–2535.

Asparouhov, T., & Muthén, B. (2008). Multilevel mixture models. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in Latent Variable Mixture Models* (pp. 27–51). Information Age Publishing, Inc.

Asparouhov, T., & Muthén, B. (2016). General random effect latent variable modeling: Random subjects, items, contexts, and parameters. In J. R. Harring, L. M. Stapleton, & S. N. Beretvas (Eds.), *Advances in multilevel modeling for educational research: Addressing practical issues found in real-world applications* (pp. 163-192). Charlotte, NC: Information Age Publishing, Inc.

Asparouhov, T., & Muthén, B. (2021). Advances in Bayesian model fit evaluation for structural

equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(1), 1–14. https://doi.org/10.1080/10705511.2020.1764360

Bailey, P., Emad, A., Huo, H., Lee, M., Liao, Y., Lishinski, A., Nguyen, T., Xie, Q., Yu, J., Zhang, T., Buehler, E., Lee, S., Sikali, E., Bundsgaard, J., C'deBaca, R., & Christensen, A. A. (2023). EdSurvey: Analysis of NCES education survey and assessment data (R package version 3.0.2) [Computer software]. https://CRAN.R-project.org/package=EdSurvey

Bainter, S. A. (2015). Abstract: Advantages of a Bayesian approach for examining class structure in finite mixture models. *Multivariate Behavioral Research*, *50*(1), 127–127. https://doi.org/10.1080/00273171.2014.988987

Barnard, J., McCulloch, R., & Meng, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, *10*(4), 1281–1311.

Bauer, D. J. (2003). Estimating multilevel linear models as structural equation models. *Journal of Educational and Behavioral Statistics*, *28*(2), 135–167. https://doi.org/10.3102/10769986028002135

Blozis, S. A. (2022). A latent variable mixed-effects location scale model with an application to daily diary data. *Psychometrika*, *87*(4), 1548–1570. https://doi.org/10.1007/s11336-022-09864-8

Bollen, K. A., & Stine, R. A. (1992). Bootstrapping goodness-of-fit measures in structural equation models. *Sociological Methods & Research*, *21*(2), 205–229. https://doi.org/10.1177/0049124192021002004

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker,

M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1). https://doi.org/10.18637/jss.v076.i01

Cassiday, K. R., Cho, Y., & Harring, J. R. (2021). A comparison of label switching algorithms in the context of growth mixture models. *Educational and Psychological Measurement*, *81*(4), 668–697. https://doi.org/10.1177/0013164420970614

Cleveland, W. S., Denby, L., & Liu, C. (2002). Random location and scale effects: Model building methods for a general class of models. *Computing Science and Statistics*, *32*, 3–10.

Cole, D. A., & Preacher, K. J. (2014). Manifest variable path analysis: Potentially serious and misleading consequences due to uncorrected measurement error. *Psychological Methods*, *19*(2), 300–315. https://doi.org/10.1037/a0033805

Cox, T. H., & Blake, S. (1991). Managing cultural diversity: Implications for organizational competitiveness. *Academy of Management Perspectives*, *5*(3), 45–56. https://doi.org/10.5465/ame.1991.4274465

Curran, P. J. (2003). Have multilevel models been structural equation models all along? *Multivariate Behavioral Research*, *38*(4), 529–569. https://doi.org/10.1207/s15327906mbr3804_5

De Roover, K., Vermunt, J. K., & Ceulemans, E. (2020). Mixture multigroup factor analysis for unraveling factor loading noninvariance across many groups. *Psychological Methods*, *27*(3), 281–306. https://doi.org/10.1037/met0000355

Deboeck, P. R., Montpetit, M. A., Bergeman, C. S., & Boker, S. M. (2009). Using derivative estimates to describe intraindividual variability at multiple time scales. *Psychological Methods*, *14*(4), 367–386. https://doi.org/10.1037/a0016622

Decristan, J., Fauth, B., Kunter, M., Büttner, G., & Klieme, E. (2017). The interplay between class heterogeneity and teaching quality in primary school. *International Journal of Educational Research*, *86*, 109–121. https://doi.org/10.1016/j.ijer.2017.09.004

Depaoli, S. (2013). Mixture class recovery in GMM under varying degrees of class separation: Frequentist versus Bayesian estimation. *Psychological Methods*, *18*(2), 186–219. https://doi.org/10.1037/a0031609

Depaoli, S., & Clifton, J. P. (2015). A Bayesian approach to multilevel structural equation modeling with continuous and dichotomous outcomes. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*(3), 327–351. https://doi.org/10.1080/10705511.2014.937849

Dzubur, E., Ponnada, A., Nordgren, R., Yang, C.-H., Intille, S., Dunton, G., & Hedeker, D. (2020). MixWILD: A program for examining the effects of variance and slope of time-varying variables in intensive longitudinal data. *Behavior Research Methods*, *52*(4), 1403–1427. https://doi.org/10.3758/s13428-019-01322-1

Easley, C. A. (2001). Developing valuing and managing diversity in the new millennium. *Organization Development Journal, 19*(4), 38–51.

Fagot, D., Mella, N., Borella, E., Ghisletta, P., Lecerf, T., & De Ribaupierre, A. (2018). Intra-Individual variability from a lifespan perspective: a comparison of latency and accuracy measures. *Journal of Intelligence*, *6*(1), 16–33. https://doi.org/10.3390/jintelligence6010016

Feng, Y., & Hancock, G. R. (2020). Abstract: An analytical framework for research questions with variability as the outcome or predictor. *Multivariate Behavioral Research*, *55*(1), 144–145. https://doi.org/10.1080/00273171.2019.1695570

Feng, Y., & Hancock, G. R. (2022). A structural equation modeling approach for modeling variability as a latent variable. *Psychological Methods*, *Advance online publication*. https://doi.org/10.1037/met0000477

Garnier-Villarreal, M., & Jorgensen, T. D. (2020). Adapting fit indices for Bayesian structural equation modeling: Comparison to maximum likelihood. *Psychological Methods*, *25*(1), 46–70. https://doi.org/10.1037/met0000224

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculations of posterior moments. *Bayesian Statistics*, *4*, 641–649.

Goldstein, H. (2011). *Multilevel Statistical Models* (4th ed.). John Wiley & Sons.

Goldstein, H., & McDonald, R. P. (1988). A general model for the analysis of multilevel data. *Psychometrika*, *53*(4), 455–467. https://doi.org/10.1007/BF02294400

Hamaker. E. L. (2012). Why researchers should think "within-person": A paradigmatic rationale. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 43–61). Guilford Press.

Heck, R. H., & Reid, T. (2023). Multilevel structural equation modeling. In R. Hoyle (Ed.), *Handbook of structural equation modeling* (2nd ed.). Guilford Press.

Hedeker, D., & Mermelstein, R. J. (2007). Mixed-effects regression models with heterogeneous variance: Analyzing ecological momentary assessment (EMA) data of smoking. In T. D. Little, J. A. Bovaird, & N. A. Card (Eds.), *Modeling contextual effects in longitudinal studies* (pp. 183–206). Lawrence Erlbaum Associates Publishers.

Hedeker, D., & Mermelstein, R. J. (2012). Mood changes associated with smoking in adolescents: An application of a mixed-effects location scale model for longitudinal ecological momentary assessment data. In G. R. Hancock & J. R. Harring (Eds.),

Advances in longitudinal methods in the social and behavioral sciences (pp. 59–79). IAP Information Age Publishing.

Hedeker, D., Mermelstein, R. J., & Demirtas, H. (2008). An application of a mixed-effects location scale model for analysis of ecological momentary assessment (EMA) data. *Biometrics*, *64*(2), 627–634. https://doi.org/10.1111/j.1541-0420.2007.00924.x

Hedeker, D., Mermelstein, R. J., & Demirtas, H. (2012). Modeling between-subject and within-subject variances in ecological momentary assessment data using mixed-effects location scale models. *Statistics in Medicine*, *31*(27), 3328–3336. https://doi.org/10.1002/sim.5338

Hedeker, D., & Nordgren, R. (2013). MIXREGLS: A program for mixed-effects location scale analysis. *Journal of Statistical Software*, *52*(12). https://doi.org/10.18637/jss.v052.i12

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, *29*(1), 60–87. https://doi.org/10.3102/0162373707299706

Hoofs, H., van de Schoot, R., Jansen, N. W. H., & Kant, I. (2018). Evaluating model fit in bayesian confirmatory factor analysis with large samples: simulation study introducing the BRMSEA. *Educational and Psychological Measurement*, *78*(4), 537–568. https://doi.org/10.1177/0013164417709314

Hox, J., Moerbeek, M., & van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications (3rd ed.). Routledge. https://doi.org/10.4324/9781315650982*

Hoyle, R. H., & Gottfredson, N. C. (2015). Sample size considerations in prevention research applications of multilevel modeling and structural equation modeling. *Prevention Science*, *16*(7), 987–996. https://doi.org/10.1007/s11121-014-0489-8

Hultsch, D. F., & MacDonald, S. W. S. (2004). Intraindividual variability in performance as a theoretical window onto cognitive aging. In R. Dixon, L. Backman, & L.-G. Nilsson (Eds.), *New Frontiers in Cognitive Aging* (pp. 65–88). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198525691.003.0004

Jak, S. (2014). Testing strong factorial invariance using three-level structural equation modeling. *Frontiers in Psychology*, *5*. https://www.frontiersin.org/articles/10.3389/fpsyg.2014.00745

Jak, S. (2019). Cross-level invariance in multilevel factor models. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(4), 607–622.

Jak, S., & Jorgensen, T. D. (2017). Relating measurement invariance, cross-level invariance, and multilevel reliability. *Frontiers in Psychology*, *8*. https://www.frontiersin.org/articles/10.3389/fpsyg.2017.01640

Jak, S., Jorgensen, T. D., & Rosseel, Y. (2021). Evaluating cluster-level factor models with lavaan and Mplus. *Psych*, *3*(2), 134–152. https://doi.org/10.3390/psych3020012

Jak, S., Oort, F. J., & Dolan, C. V. (2013). A test for cluster bias: detecting violations of measurement invariance across clusters in multilevel data. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*(2), 265–282. https://doi.org/10.1080/10705511.2013.769392

Jehn, K. A., Northcraft, G. B., & Neale, M. A. (1999). Why differences make a difference: a field study of diversity, conflict and performance in workgroups. *Administrative Science Quarterly*, *44*(4), 741–763. https://doi.org/10.2307/2667054

Jongerling, J., Laurenceau, J.-P., & Hamaker, E. L. (2015). A multilevel AR(1) model: Allowing for inter-individual differences in trait-scores, inertia, and innovation variance.

Multivariate Behavioral Research, *50*(3), 334–349. https://doi.org/10.1080/00273171.2014.1003772

Kapur, K., Li, X., Blood, E. A., & Hedeker, D. (2015). Bayesian mixed-effects location and scale models for multivariate longitudinal outcomes: An application to ecological momentary assessment data. *Statistics in Medicine*, *34*(4), 630–651. https://doi.org/10.1002/sim.6345

Kasim, R. M., & Raudenbush, S. W. (1998). Application of Gibbs sampling to nested variance components models with heterogeneous within-group variance. *Journal of Educational and Behavioral Statistics*, *23*(2), 93–116. https://doi.org/10.3102/10769986023002093

Kaufhold, S. P., & van Leeuwen, E. J. C. (2019). Why intergroup variation matters for understanding behaviour. *Biology Letters*, *15*(11), 20190695. https://doi.org/10.1098/rsbl.2019.0695

Kim, E. S., Cao, C., Wang, Y., & Nguyen, D. T. (2017). Measurement invariance testing with many groups: a comparison of five approaches. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(4), 524–544. https://doi.org/10.1080/10705511.2017.1304822

Kim, E. S., Joo, S.-H., Lee, P., Wang, Y., & Stark, S. (2016). Measurement invariance testing across between-level latent classes using multilevel factor mixture modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(6), 870–887.

Kim, E. S., & 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. *Frontiers in Psychology*, *8*. https://www.frontiersin.org/articles/10.3389/fpsyg.2017.01499

Kim, E. S., & Wang, Y. (2019). Investigating sources of heterogeneity with three-step multilevel

factor mixture modeling: Beyond testing measurement invariance in cross-national studies. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(2), 165–181. https://doi.org/10.1080/10705511.2018.1521284

Kim, E. S., Wang, Y., & Kiefer, S. M. (2018). Cross-level group measurement invariance when groups are at different levels of multilevel data. *Educational and Psychological Measurement*, *78*(6), 973–997. https://doi.org/10.1177/0013164417739062

Knebel, D., Ayali, A., Guershon, M., & Ariel, G. (2019). Intra- versus intergroup variance in collective behavior. *Science Advances*, *5*(1), eaav0695. https://doi.org/10.1126/sciadv.aav0695

Kupers, E., Lehmann-Wermser, A., McPherson, G., & van Geert, P. (2019). Children's creativity: A theoretical framework and systematic review. *Review of Educational Research*, *89*(1), 93–124. https://doi.org/10.3102/0034654318815707

Leckie, G., French, R., Charlton, C., & Browne, W. (2014). Modeling heterogeneous variance–covariance components in two-level models. *Journal of Educational and Behavioral Statistics*, *39*(5), 307–332. https://doi.org/10.3102/1076998614546494

Lee, Y., & Nelder, J. A. (2001). Hierarchical generalised linear models: A synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika*, *88*(4), 987–1006. https://doi.org/10.1093/biomet/88.4.987

Lee, Y., & Nelder, J. A. (2006). Double hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *55*(2), 139–185. https://doi.org/10.1111/j.1467-9876.2006.00538.x

Lee, Y., Ronnegard, L., & Noh, M. (2017). *Data analysis using hierarchical generalized linear models with R*. CRC Press.

Levy, R. (2011). Bayesian data-model fit assessment for structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *18*(4), 663–685. https://doi.org/10.1080/10705511.2011.607723

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. https://doi.org/10.1111/j.2044-8317.1989.tb00911.x

McNeish, D. (2021). Specifying Location-Scale Models for Heterogeneous Variances as Multilevel SEMs. *Organizational Research Methods*, *24*(3), 630–653. https://doi.org/10.1177/1094428120913083

McNeish, D., & Hamaker, E. L. (2020). A primer on two-level dynamic structural equation models for intensive longitudinal data in Mplus. *Psychological Methods*, *25*(5), 610–635. https://doi.org/10.1037/met0000250

McNeish, D. M., & Stapleton, L. M. (2016). The effect of small sample size on two-level model estimates: A review and illustration. *Educational Psychology Review*, *28*(2), 295–314. https://doi.org/10.1007/s10648-014-9287-x

McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods*, *52*(6), 2287–2305. https://doi.org/10.3758/s13428-020-01398-0

Mumme, R., Pushpanathan, M., Donaldson, S., Weinborn, M., Rainey-Smith, S. R., Maruff, P., & Bucks, R. S. (2021). Longitudinal association of intraindividual variability with cognitive decline and dementia: A meta-analysis. *Neuropsychology*, *35*(7), 669–678. https://doi.org/10.1037/neu0000746

Muthén, B., & Asparouhov, T. (2009). Multilevel regression mixture analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *172*(3), 639–657.

https://doi.org/10.1111/j.1467-985X.2009.00589.x

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*(4), 557–585. https://doi.org/10.1007/BF02296397

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research*, *22*(3), 376–398. https://doi.org/10.1177/0049124194022003006

Muthén, L. K., & Muthén, B. O. (1998-2023). *Mplus user's guide* (8th ed.). Muthén & Muthén.

Nagelkerke, E., Oberski, D. L., & Vermunt, J. K. (2016). Goodness-of-fit of multilevel latent class models for categorical data. *Sociological Methodology*, *46*(1), 252–282. https://doi.org/10.1177/0081175015581379

Nelder, J. A., & Lee, Y. (1991). Generalized linear models for the analysis of taguchi-type experiments. *Applied Stochastic Models and Data Analysis*, *7*(1), 107–120. https://doi.org/10.1002/asm.3150070110

Nesselroade, J. R. (1991). Interindividual differences in intraindividual change. In Collins, L. M., & Horn, J. L. (Eds.), *Best methods for the analysis of change: Recent advances, unanswered questions, future directions* (pp. 92–105). American Psychological Association. https://doi.org/10.1037/10099-006

Nesselroade, J. R., & Ram, N. (2004). Studying intraindividual variability: What we have learned that will help us understand lives in context. *Research in Human Development*, *1*(1–2), 9–29. https://doi.org/10.1207/s15427617rhd0101&2_3

Nestler, S. (2020). Modelling inter-individual differences in latent within-person variation: The confirmatory factor level variability model. *British Journal of Mathematical and Statistical Psychology*, *73*(3), 452–473. https://doi.org/10.1111/bmsp.12196

Nestler, S. (2021). Modeling intraindividual variability in growth with measurement burst

designs. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(1), 28–39. https://doi.org/10.1080/10705511.2020.1757455

Nylund-Gibson, K., Grimm, R. P., & Masyn, K. E. (2019). Prediction from latent classes: A demonstration of different approaches to include distal outcomes in mixture models. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(6), 967–985. https://doi.org/10.1080/10705511.2019.1590146

Nylund-Gibson, K., & Masyn, K. (2016). Covariates and mixture modeling: Results of a simulation study exploring the impact of misspecified effects on class enumeration. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*, 1–16. https://doi.org/10.1080/10705511.2016.1221313

OECD. (2013a). *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy.* OECD Publishing. Https://doi.org/10.1787/9789264190511.

OECD. (2013b). *PISA 2012 dataset* [Data file and code book]. Https:// www.oecd.org/pisa/data/pisa2012database-downloadabledata.htm.

OECD. (2013c). *PISA 2012 technical report.* OECD Publishing. Https:// www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf.

Papastamoulis, P. (2016). label.switching: An R package for dealing with the label switching problem in MCMC outputs [Code Snippet]. *Journal of Statistical Software*, *69*(1), 1–24. https://doi.org/10.18637/jss.v069.c01

Papastamoulis, P., & Iliopoulos, G. (2010). An artificial allocations based solution to the label switching problem in bayesian analysis of mixtures of distributions. *Journal of Computational and Graphical Statistics*, *19*(2), 313–331.

https://doi.org/10.1198/jcgs.2010.09008

Petersen, L.-E., Dietz, J., & Frey, D. (2004). The effects of intragroup interaction and cohesion on intergroup bias. *Group Processes & Intergroup Relations*, *7*(2), 107–118. https://doi.org/10.1177/1368430204041397

Plummer, M., Stukalov, A., & Denwood, M. (2022). *Bayesian graphical models using MCMC* (R package version 4-13) [Computer software]. https://cran.r-project.org/web/packages/rjags/rjags.pdf

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika*, *69*(2), 167–190. https://doi.org/10.1007/BF02295939

Rabe-Hesketh, S., Skrondal, A., & Zheng, X. (2012). Multilevel structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 512–531). Guilford Press.

Raitano, M., & Vona, F. (2013). Peer heterogeneity, school tracking and students' performances: Evidence from PISA 2006. *Applied Economics*, *45*(32), 4516–4532. https://doi.org/10.1080/00036846.2013.791020

Ram, N., Gatzke-Kopp, L., Gerstorf, D., Coccia, M., Morack, J., & Molenaar, P. C. M. (2015). Intraindividual variability across the life span: Moving toward computational developmental science. In M. Diehl, K. Hooker, & M. Sliwinski (Eds.), *Handbook of intraindividual variability across the life span* (pp. 16–34). Routledge/Taylor & Francis Group.

Rast, P., & Ferrer, E. (2018). A mixed-effects location scale model for dyadic interactions. *Multivariate Behavioral Research*, *53*(5), 756–775. https://doi.org/10.1080/00273171.2018.1477577

Rast, P., Hofer, S. M., & Sparks, C. (2012). Modeling individual differences in within-person variation of negative and positive affect in a mixed effects location scale model using BUGS/JAGS. *Multivariate Behavioral Research*, *47*(2), 177–200. https://doi.org/10.1080/00273171.2012.658328

Raudenbush, S. W., & Bryk, A. S. (1987). Examining correlates of diversity. *Journal of Educational Statistics*, *12*(3), 241–269. https://doi.org/10.3102/10769986012003241

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed). Sage Publications. http://catdir.loc.gov/catdir/enhancements/fy0658/2001057880-t.html

Redner, R. A., & Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM Algorithm. *SIAM Review*, *26*(2), 195–239. https://doi.org/10.1137/1026034

Rodríguez, C. E., & Walker, S. G. (2014). Label switching in Bayesian mixture models: Deterministic relabeling strategies. *Journal of Computational and Graphical Statistics*, *23*(1), 25–45. https://doi.org/10.1080/10618600.2012.735624

Ryu, E. (2014). Model fit evaluation in multilevel structural equation models. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00081

Ryu, E., & West, S. G. (2009). Level-specific evaluation of model fit in multilevel structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(4), 583–601. https://doi.org/10.1080/10705510903203466

Schultzberg, M., & Muthén, B. (2018). Number of subjects and time points needed for multilevel time-series analysis: A simulation study of dynamic structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(4), 495–515. https://doi.org/10.1080/10705511.2017.1392862

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, *52*(3), 333–343. https://doi.org/10.1007/BF02294360

Snijders, T. A. B., & Bosker, R. J. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. SAGE.

Stapleton, L. M., Yang, J. S., & Hancock, G. R. (2016). Construct meaning in multilevel settings. *Journal of Educational and Behavioral Statistics*, *41*(5), 481–520. https://doi.org/10.3102/1076998616646200

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *62*(4), 795–809. https://doi.org/10.1111/1467-9868.00265

Sterba, S. K., & Rights, J. D. (2022). R-squared measures for multilevel mixture models with random effects. *Structural Equation Modeling: A Multidisciplinary Journal*, *29*(4), 489–506. https://doi.org/10.1080/10705511.2021.1962325

Tofighi, D., & Enders, C. K. (2008). Identifying the Correct Number of Classes in a Growth Mixture Model. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in Latent Variable Mixture Models* (pp. 317–341). Information Age.

Tueller, S., & Lubke, G. (2010). Evaluation of structural equation mixture models: Parameter estimates and correct class assignment. *Structural Equation Modeling: A Multidisciplinary Journal*, *17*(2), 165–192. https://doi.org/10.1080/10705511003659318

van Dijk, M., & van Geert, P. (2015). The nature and meaning of intraindividual variability in development in the early life span. In M. Diehl, K. Hooker, & M. J. Sliwinski (Eds.), *Handbook of intraindividual variability across the life span* (pp. 37–58). Routledge/Taylor & Francis Group.

van Knippenberg, D., De Dreu, C. K. W., & Homan, A. C. (2004). Work group diversity and group performance: An integrative model and research agenda. *Journal of Applied Psychology*, *89*(6), 1008–1022. https://doi.org/10.1037/0021-9010.89.6.1008

Vermunt, J. K. (2008). Latent class and finite mixture models for multilevel data sets. *Statistical Methods in Medical Research*, *17*(1), 33–51.

Vermunt, J. K. (2011). Mixture models for multilevel data sets. In J. J. Hox & J. K. Roberts (Eds.). *Handbook of advanced multilevel analysis* (pp. 59–81). Routledge.

Vermunt, J. K., & Magidson, J. (2005). Hierarchical mixture models for nested data structures. In C. Weihs & W. Gaul (Eds.), *Classification—The ubiquitous challenge* (pp. 240–247). Springer. https://doi.org/10.1007/3-540-28084-7_26

Wang, L. P., Hamaker, E., & Bergeman, C. S. (2012). Investigating inter-individual differences in short-term intra-individual variability. *Psychological Methods*, *17*(4), 567–581. https://doi.org/10.1037/a0029317

Williams, D. R., Mulder, J., Rouder, J. N., & Rast, P. (2021). Beneath the surface: Unearthing within-person variability and mean relations with Bayesian mixed models. *Psychological Methods*, *26*, 74–89. https://doi.org/10.1037/met0000270