University of California

Los Angeles

Estimation of Contextual Effects through Nonlinear Multilevel Latent Variable Modeling with a Metropolis-Hastings Robbins-Monro Algorithm

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Education

by

Ji Seung Yang

2012

Abstract of the Dissertation

Estimation of Contextual Effects through Nonlinear Multilevel Latent Variable Modeling with a Metropolis-Hastings Robbins-Monro Algorithm

by

Ji Seung Yang

Doctor of Philosophy in Education

University of California, Los Angeles, 2012

Professor Li Cai, Chair

Nonlinear multilevel latent variable modeling has been suggested as an alternative to traditional hierarchical linear modeling to more properly handle measurement error and sampling error issues in contextual effects modeling. However, a nonlinear multilevel latent variable model requires significant computational effort because the estimation process involves high dimensional numerical integration, particularly when the number of latent variables is large. The main purpose of this study is to improve estimation efficiency in obtaining full-information maximum likelihood (FIML) estimates of contextual effects by adopting the Metropolis-Hastings Robbins-Monro algorithm (MH-RM; Cai, 2008, 2010a, 2010b). This study considers contextual effects not only as compositional effects but also as cross-level interactions, in which latent variables are measured by categorical manifest variables. R programs (R Core Team, 2012) implementing the MH-RM algorithm were produced to fit nonlinear multilevel latent variable models. Computational efficiency and parameter recovery were assessed by comparing results with an EM algorithm that uses adaptive Gauss-Hermite quadrature for numerical integration. Results indicate that the MH-RM algorithm can produce FIML estimates and their standard errors efficiently, and the efficiency of MH-RM was more prominent for a cross-level interaction model, which requires 5-dimensional integration. Simulations, with various sampling and measurement structure conditions, were conducted to obtain information about the performance of nonlinear multilevel latent variable modeling compared to traditional hierarchical linear modeling. Results suggest that nonlinear multilevel latent variable modeling can more properly estimate and detect a contextual effect and a cross-level interaction than the traditional approach. As empirical illustrations, two subsets of data extracted from Programme for International Student Assessment (PISA; OECD, 2000) were used. A negative contextual effect was found from the U.S. data in terms of the relationship between reading literacy and self-concept about reading, supporting results from previous studies. A negative, but not statistically significant, cross-level interaction was found between reading literacy and co-operative learning preference from the analysis of data collected in Korea.

The dissertation of Ji Seung Yang is approved.

Sandra Graham

Steven Paul Riese

Michael Seltzer

Li Cai, Committee Chair

University of California, Los Angeles

2012

To my parents

Table of Contents

List of Figures

List of Tables

Acknowledgments

I believe that this dissertation is not an end but only a beginning. Still, to get to this starting point, I have been indebted to countless individuals for their support. I am truly blessed to express my gratitude on this page.

First and foremost, I would like to thank my advisors, Drs. Li Cai and Michael Seltzer for all the support, guidance, assistance, and encouragement over the years. I am always and deeply grateful to have them as my advisors and mentors who have moved me with their admirable intelligence and genuine hearts toward not only academic research but also people. Particularly, this dissertation would not have been possible without Dr. Cai's devotion from beginning to the end. I could never thank him enough and I deeply appreciate his being in my academic life. I would also like to thank the members of my committee, Drs. Sandra Graham and Steve Riese, for their stimulating discussions and suggestions. In addition, I would like to thank my advisors at Yonsei Univeristy, Drs. Sangjin Kang and Guemin Lee, for their mentoring and guidance since I entered the fields of educational measurement and quantitative research methodology.

I would like to acknowledge the academic lineage of my former and current advisors: Drs. Anthony Bryk (Michael Seltzer's advisor), Stephen Raudenbush (Sangjin Kang's advisor), David Thissen and Robert MacCallum (Li Cai's advisors), David Frisbie and Robert Brennan (Guemin Lee's advisor). I am very honored to be able to even mention them here through direct and indirect connections, and I am aware that major concepts and thoughts in this dissertation have been cast by these great scholars and developed through the privileged academic lineage. I particularly express my deepest gratitude toward Drs. David Thissen and Robert Brennan for taking care of and encouraging me whenever we meet at conferences.

I also want to express my appreciation of the social research methodology program, a wonderful community with a highly supportive culture. I have been taught and supported intellectually and emotionally by Pfs. Noreen Web, Mike Rose, and José Felipe Martínez Fernández, both in and out of the classroom, and I deeply appreciate their

guidance and encouragement. I would also like to thank Drs. Eva Baker and Kilchan Choi for their support and guidance while I was working at CRESST. I thank the incredible colleagues in Dr. Cai's psychometric lab, specifically Taehun Lee, Mark Hansen, Scott Monroe, Larry Thomas, and Moonsoo Lee, for their intellectual and emotional support and help. Mark and Scott deserve special thanks for all the help they have given. I cannot fail to mention the following great peers for their support and encouragement over the years: Drs. (or soon-to-be-doctors) Jordan Rickles, Anne Vo, Regina Richter, Cecilia Henríquez Fernández, Deirdre Kerr, Sandra Naranjo, Alejandra Priede Schubert, Lisa Dillman, Belinda Thompson, Nami Shin, Rebecca Luskin and Meiko Lin. Anne deserves special thanks for editing my manuscript and feeding me while I was writing it. I would like to extend my thanks to the office staff for their help, particularly Amy Gershon, Kim Mattheussens, and Matthew Palmer. I also want to thank Dr. Hyekyung Jung, Dr. Cheong Huh, Dr. Youngsoon So, Dr. Jinok Kim and Yourim Chai for encouraging and motivating me in numerous ways as international students who walked this path before me. I should acknowledge that their being in LA itself was a relief.

While spending years in LA, I have been supported by not only the UCLA community but also multiple groups of mentors, colleagues, and friends outside of UCLA. I would like to thank Drs. Myungae Lee, Kyungseok Min, Minyoung Song, Yangboon Kim, Jooah Kim, Jihyun Jang, Yongkyum Kim, Jongmin Kim, and Yongsang Lee for their unwavering support and caring whenever and wherever I met them. Another group of people who deserve my deepest gratitude is Dr. Bright Rho, Hyesuk Jang, Hyeyun Kim, Dr. Jinnie Choi, Yeonkyung Kim, Jeongeun Lee, Miae Kang, Dr. Minjeong Jeon, and Woori Kim, who have shared tears and laughs together. I also would like to thank the Yonsei community in LA, particularly Sunyoon Kim, Sophia Yoo's family, Dr. Hyunjoo Hwang, David Lee, and Hyeyoung Jung for their being here and literally taking care of me since we met.

I acknowledge that I have benefited from LaTex and R Group. The Edwin and Edna C. Lee Fellowship and multiple travel grants through the department of education and Yonsei International Foundation Scholarship (2011) are also acknowledged. Financial support through grants from the Institute of Education Sciences (R305B080016 and R305D100039), the National Institute on Drug Abuse (R01DA026943 and R01DA030466), and Society of Multivariate Experimental Psychology (dissertation support awards) are greatly appreciated. The views expressed in this paper are mine alone and do not reflect the views/policies of the funding agencies or grantees.

Finally, I thank my family for their immeasurable love - my beloved parents, my only brother Jungwon, his wife Yunha, my sweet aunt Yeonsil, and uncle Dr. Changshik Yang's family that includes my incredible cousins, Hank and his siblings. Hank also deserves special thanks for providing me with occasional and urgent editing as well as heartwarming emotional support when it is needed. I also deeply appreciate the man who has inspired and influenced these three families and passed away while I was working on this manuscript. I ultimately thank my God, who has guided me to meet all the people mentioned above and to learn knowledge, life, and love, through them.

Vita

2002	A.A., General Studies Montgomery college Rockville, MD
2005	B.A., Education College of Sciences in Education Yonsei University Seoul, Korea
2007	M.A., Education (Educational measurement and evaluation) College of Sciences in Education Yonsei University Seoul, Korea
2005–2006	Teaching Assistant and Research Assistant Department of Education Yonsei University Seoul, Korea
2006-2007	Appointed Researcher Korea Institute of Curriculum and Evaluation (KICE) Seoul, Korea
2007-2010	Graduate Student Researcher National Center for Research on Evaluation, Standards, and Student Testing (CRESST) University of California, Los Angeles Los Angeles, CA

2010–2012	Special Reader and Graduate Student Researcher Department of Education University of California, Los Angeles Los Angeles, CA

Publications

Yang, J. S., & Lee, G. (2007). Estimating reliability of test scores composed of testlets using generalizability theory approaches. *Korean Journal of Educational Evaluation, 20(1)*, 119-139.

Yang, J. S., Lee, G., & Kang, S. (2007). Estimating reliability of test scores composed of testlets using item response theory approaches. *Korean Journal of Educational Evaluation, 20(3)*, 147-167.

Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods, 16(3)*, 221-248.

Yang, J. S., Hansen, M., & Cai, L. (2012). Characterizing sources of uncertainty in item response theory scale scores. *Educational and Psychological Measurement, 72(2)*, 264-290.

CHAPTER 1

Introduction

This study adopts a Metropolis-Hastings Robbins-Monro (MH-RM; Cai, 2008, 2010a, 2010b) algorithm to estimate contextual effects more efficiently in the *multilevel latent variable modeling* framework. This chapter provides a review of relevant background and research goals. The following section discusses how to define and specify contextual effects in a statistical model and the methodological issues that have drawn researchers' attention.

1.1 Background

1.1.1 Contextual Effects

The Greek philosopher Aristotle said, "Man is a social animal. [...] He who lives without society is either a beast or a God" in *Politics I.2*. Since human beings are social, their behaviors are naturally influenced by social groups such as one's family, classroom, school, workplace, and country. The study of the roles of group context on actions and attitudes of individuals is called *contextual analysis* (Iversen, 1991). Without the influence of social context, any individual-level relationship between two variables of interest will be constant across groups, meaning the group-level relationship is the same as the individual-level relationship. In this case, two individuals who have the same characteristics are expected to have the same outcome.

However, it is possible for two individuals to have different outcome levels even though their individual characteristics are the same. When the difference in expected outcomes can be explained by a group-level variable, we take the difference as an effect of social context. By decomposing the effect of a predictor on an outcome at both the individual and group levels, the effect of social context can be investigated. Accordingly, a *contextual effect* or a *compositional effect* is defined as the extent to which the magnitude of the group-level relationship differs from the individual-level effects (see, e.g., Raudenbush & Bryk, 2002). Understanding human behaviors through not only an individual level perspective, but also the lens of social context, helps social science researchers obtain a more complete picture of individuals as well as society. Therefore, methodologists have tried to quantify contextual effects using statistical models that are known as *contextual models*.

An interesting aspect of the relationship between individuals and groups is that the direction of influence is not unilateral. Group context affects individual actions or attitudes, but the group context is often formed from individual characteristics. In many cases, the interaction between an organization and an individual leads to a contextual effect. The rationale of using a cross-level interaction term in statistical modeling for a contextual effect lies in the interactive dynamics between individuals and groups.

The particular contextual effect of interest in this study is one that occurs when a group-level characteristic of interest is measured by individual-level characteristics, which is different from the case where a group-level characteristic is simply defined as deterministic categories such as a public school or a private school in an educational setting. Lüdtke et al. (2008) discussed two different aggregation processes in constructing a group-level construct by aggregating individual data at the group level: reflective and formative. The former assumes an "isomorphic relationship" between the individual-level data and the group-level construct, while the latter assumes the group-level variable is a simple index of level-1 construct. This study considers the reflective aggregation in general. However, formative aggregation can also benefit from this multilevel latent variable modeling in that possible measurement error in level-1 can be considered. The modeling also provides opportunities for further qualitative research. For example, a composition of students in terms of gender or ethnicity can be a contextual variable. However, further analysis to investigate what kind of cultural components or psychological constructs account for differences in the phenomena will eventually require reflective aggregation.

Contextual models are widely applied in organizational and industrial psychology, where researchers try to isolate the effects of individual persons from those of larger groups (see, e.g., Firebaugh, 1978; Erbring & Young, 1979). Iversen (1991) pointed out that educational research has been "the major user" of contextual analysis among social science disciplines. This is because of: 1) the nature of typical educational data in which students are nested in schools, and 2) the nature of educational research questions that connect school-level or class-level characteristics to student-level outcomes. Educational researchers' endeavor to open the *black box* between school input and outcome started more actively since Coleman, Hoffer, and Kilgore (1982) reported skeptical results about school effects. Therefore, most previous contextual analyses cannot be separated from the proliferation of *hierarchical linear models* (HLM [1], Raudenbush & Bryk, 2002).

In educational research, a contextual effect has been traditionally defined as the difference between two coefficients in the multilevel analysis framework (Raudenbush & Bryk, 1986; Willms, 1986; Lee & Bryk, 1989; Raudenbush & Willms, 1995): one from the individual-level and the other coefficient from the school-level. A representative application of this kind of contextual effect in education is discussed in Raudenbush and Bryk (2002) using a subset of High School and Beyond Data (HS&B). In this example, individual math achievement is regressed on individual-level socioeconomic status (SES) and school-level math achievement is regressed on aggregated school-level SES using multilevel modeling. The result shows that two coefficient estimates are not the same, indicating two students who have the same SES level are expected to have different levels of math achievement depending on to which school a student belongs. Statistically significant difference between these two coefficients represents a significant compositional effect. Moreover, the contextual effect in the study was interpreted as the positive

[1] HLM has various names such as *multilevel linear models* in sociological research (Mason, Wong, & Entwisle, 1983), *mixed-effects models* or *random effect models* in biometric research (Elston & Grizzle, 1962; Laird & Ware, 1982; Singer, 1998). *Random-coefficient regression models* (Rosenberg, 1973; Longford, 1993) and *covariance components models* (Dempster, Rubin, & Tsutakawa, 1981; Longford, 1987) also refer to the same kind of models in econometrics and statistical literature, respectively.

increment to student-level learning by virtue of attending a school which has higher school-level SES (see. Figure 1.1).

Another example of a contextual effect in psychology is the "bigfish-little-pond effect (BFLPE)" in which the magnitude of effect of student-level achievement on student academic self-concept is not consistent across classrooms or schools (Marsh, 1987). More precisely, the relationship between student-level achievement and student academic self-concept is positive, but the magnitude of the effect at school-level is found to be different from the effect of achievement on academic self-concept at student-level. Accordingly, two students who have the same academic achievement can have different levels of academic self-concept depending on the school achievement levels in which each student is situated. In this case, the student who belongs to a school with lower average achievement shows greater self-concept just like a fish that feels it is big because it is in a little pond. More applications of contextual models can be found in organizational research (see, e.g., Bliese, 2000; Kozlowski & Klein, 2000; Bliese, Chan, & Ployhart, 2007; LaHuis & Ferguson, 2009).

Contextual effect models are widely applied in other social science disciplines such as criminology (e.g., Bottoms & Wiles, 2004; Wikström, 1998; Wooldredge & Thistlethwaite, 1999; Sampson, Morenoff, & Gannon-Rowley, 2002; Oberwittler, 2004) and public health research (e.g., Iversen, 1991; Croon & van Veldhoven, 2007; Henry & Slater, 2007). In those studies, not only schools but also other groupings, such as neighborhoods or hospitals, are considered. As an example, the contextual effect of neighborhood on serious juvenile offenses has been studied by investigating the role of subcultural values and social disorganization (Oberwittler, 2004). After controlling the effect of individual predictors, the study found that serious offences were more frequently related to adolescents with attitudes typical of delinquent subcultures and those who have lower neighborhood-level social capital, in particular. The level of social capital in Oberwittler (2004) was measured from individual survey responses just as school-level SES was measured from student-level SES.

1.1.2 Modeling Contextual Effects as Compositional Effects

As briefly mentioned in section 1.1.1, an appropriate modeling framework for contextual effects was not available before hierarchical linear models (HLM) were developed. HLM can handle nested data, properly accounting for dependence among individuals in the same level-2 unit (Raudenbush & Bryk, 2002). Therefore, contextual effect analysis has long been conducted within the HLM framework.

A traditional contextual effect model is illustrated in Figure 1.1 (Raudenbush & Bryk, 2002). In this setting, level-1 is the student level and level-2 is the school level; the predictor is SES, and the outcome is math achievement. The figure shows that the association between student-level SES and student level math achievement β_w is different from the school-level SES and school level achievement β_b, and the difference β_c is defined as the contextual effect. Therefore, β_c is the expected difference in math achievement between two students who have the same SES level but who attend different schools in terms of school-mean SES. In other words, the compositional effect β_c is the expected difference in achievement between two students who are similar in terms of family SES, but who attend schools that differ by 1 unit in their mean SES values.

The corresponding HLM can be written as follows.

$$\begin{aligned}
Y_{ij} &= \beta_{0j} + \beta_{1j}(X_{ij} - \overline{X}_{.j}) + r_{ij}, \\
\beta_{0j} &= \gamma_{00} + \gamma_{01}(\overline{X}_{.j} - \overline{X}_{..}) + u_{0j}, \\
\beta_{1j} &= \gamma_{10}, \\
\gamma_{10} &= \beta_w, \\
\gamma_{01} &= \beta_b, \\
\beta_c &= \gamma_{01} - \gamma_{10} \qquad (1.1)
\end{aligned}$$

In Equation (1.1), Y_{ij} and X_{ij} denote outcome and predictor values of student i in school j, respectively. Y_{ij} and X_{ij} are typically constructed by summing item scores on self-report responses. The random effects r_{ij} and u_{0j} are assumed to be normally distributed with zero means and variances (σ^2 and τ). In this particular definition of a *contextual effect* as a compositional effect, the within-slope, γ_{10}, is the same across groups as a fixed effect, which may or may not be appropriate, depending on the context.

1.1.3 Contextual Effects as Cross-level Interactions

In the previous compositional effect model, the within-group slopes are treated as a fixed effect, i.e., they are treated as being the same across groups. However, contextual effects can occur not only in individual outcome levels but also the individual-level outcome-predictor relationship, indicating that within-group slopes vary across groups and the contextual variable can explain some of the variance. Such a model may resemble the following:

$$\begin{aligned}
Y_{ij} &= \beta_{0j} + \beta_{1j}(X_{ij} - \overline{X}_{.j}) + r_{ij}, \\
\beta_{0j} &= \gamma_{00} + \gamma_{01}(\overline{X}_{.j} - \overline{X}_{..}) + u_{0j}, \\
\beta_{1j} &= \gamma_{10} + \gamma_{11}(\overline{X}_{.j} - \overline{X}_{..}) + u_{1j}, \\
\gamma_{10} &= \beta_w, \\
\gamma_{01} &= \beta_b, \\
\beta_c &= \gamma_{01} - \gamma_{10} \qquad (1.2)
\end{aligned}$$

The reduced form equation is:

$$\begin{aligned}
Y_{ij} &= \gamma_{00} + \gamma_{01}(\overline{X}_{.j} - \overline{X}_{..}) + \gamma_{10}(\overline{X}_{ij} - \overline{X}_{.j}) + \gamma_{11}(\overline{X}_{.j} - \overline{X}_{..})(X_{ij} - \overline{X}_{.j}) \\
&\quad + u_{1j}(X_{ij} - \overline{X}_{.j}) + u_{0j} + r_{ij} \qquad (1.3)
\end{aligned}$$

Equation (1.3) shows that the new parameter γ_{11} is the regression coefficient associated with $X_{ij}\overline{X}_{.j}$, which is a cross-level interaction term. γ_{11} captures the effect of contextual variable $\overline{X}_{.j}$ on the within-group slopes β_{1j} and eventually on the individual outcome.

As β_{1j} varies across groups, γ_{10} is the expected within-group slope when the group mean $\overline{X}_{.j}$ is the same as the grand mean $\overline{X}_{..}$. The difference between γ_{01} and γ_{10} becomes the average compositional effect, holding constant not only the influence of the individual level predictor but also the effect of the interaction between the individual-level predictor and the group-level predictor. The compositional effects here vary across groups as a function of γ_{11} and the deviation of the group-mean $\overline{X}_{.j}$ from the grand mean $\overline{X}_{..}$.

Therefore, γ_{11} captures another important aspect of compositional effect. According to Bauer and Cai (2009), omitting an interaction in HLM results in spurious large variation in slopes. Broadening the concept of contextual effects, not only as compositional effects, but also as cross-level interactions allows researchers to investigate further if the magnitude of contextual effects varies across groups.

1.1.4 Methodological Issues in Modeling Contextual effects

Though hierarchical linear modeling opened the door to estimating contextual effects, there have been two unresolved problems that have drawn researchers' attention. The first one is related to the attenuated coefficient estimates due to measurement error in predictors (Spearman, 1904), and the other is biased parameter estimates due to sampling error that are associated with aggregating level-1 variables to form level-2 variables by simply averaging the observed values. The issues are illustrated in details in the following two sections.

1.1.4.1 Measurement Error

The first source of error in estimating a contextual effect is *measurement error* using observed scores X_{ij} in modeling. Observed scores contain measurement error, unlike true scores. Measurement error is defined as the difference between a true score and an observed score in classical test theory (Allen & Yen, 2001). Borrowing some concepts from generalizability theory (Cronbach, Gleser, Nanda, & Rajaratham, 1972; Brennan, 1992), measurement error can be viewed as a type of sampling error that occurs when a limited number of items are sampled from the universe of potential items. The Item response theory framework has contributed to the definition and estimation of *standard errors of measurement* that vary across latent trait levels so that an appropriate scale can be constructed (Lord & Novick, 1968). The traditional multilevel modeling approach uses only manifest variables as both outcome and predictor variables. Therefore, those observed values are assumed to be error-free, which is not the case in most of educational research.

The effect of measurement error in predictors on regression coefficients estimates has been well known since Spearman (1904), and is typically referred to as "regression dilution" or "attenuation bias." The consequence of measurement error in multiple regression settings has been described by Fuller (1987). Raudenbush and Bryk (2002, p. 347-50) discussed similar issues in the multilevel modeling framework. More precisely, when X_{ij} is contaminated by measurement error in Equation (1.1), the estimated regression coefficient $\hat{\beta}_{1j}$ is attenuated, which leads to underadjustment of individual-level effects. In addition to $\hat{\beta}_{1j}$, $\hat{\beta}_b$ is also attenuated since $\overline{X}_{.j}$ is also contaminated by measurement error when X_{ij} is simply aggregated to level-2 to form $\overline{X}_{.j}$. Accordingly, the difference between $\hat{\beta}_b$ and $\hat{\beta}_w$ will be biased. To properly handle measurement error in predictors, researchers have paid more attention to latent variable modeling in which multiple manifest variables are used as indicators for latent variables that are free of measurement error.

Not only point estimates but also standard errors of between-level regression coefficients are expected to be underestimated when a traditional approach is taken. The dispersion matrix of estimates $\hat{\gamma}$ is,

$$Var(\hat{\gamma}) = (\sum \mathbf{W}_j^T \Delta_j^{-1} \mathbf{W}_j)^{-1}, \tag{1.4}$$

where $\mathbf{W}_j$ is a vector of between-level predictors for group j and Δ_j is the variance of $\hat{\boldsymbol{\beta}}_j$.

In HLM approach, variance of $\hat{\boldsymbol{\beta}}_j$ is defined as,

$$\begin{aligned} Var(\hat{\boldsymbol{\beta}}_j) &= Var(\mathbf{u}_j + \mathbf{e}_j) = \mathbf{T} + \mathbf{V}_j = \Delta_j \\ &= \text{parameter dispersion} + \text{error dispersion}. \end{aligned} \tag{1.5}$$

While $\mathbf{T}$ reflects the true variance of the parameters, $\mathbf{V}_j$ does not properly capture the error dispersion that comes from a measurement structure. Accordingly, smaller $\mathbf{V}_j$ yields smaller Δ_j that eventually leads underestimation of standard errors based on Equation (1.4). As statistical inferences are made based on point estimates and standard errors, the underestimated standard error issue needs to be properly addressed in modeling.

1.1.4.2 Sampling Error

The second source of error is sampling error which is associated with aggregating level-1 variables to level-2 to construct level-2 predictors. *Multi-stage probability based sampling* is often used in educational research, in which schools or districts are sampled first and classrooms or students are sampled from the upper level units. To illustrate this phenomenon, Equation (1.1) can be rewritten as a single level equation and rearranged as follows:

$$\begin{aligned} Y_{ij} &= \beta_0 + \beta_w(X_{ij} - \overline{X}_{.j}) + \beta_b(\overline{X}_{.j} - \overline{X}_{..}) + u_{0j} + r_{ij} \\ &= \beta_0 + \beta_w X_{ij} + \beta_c \overline{X}_{.j} - \beta_b \overline{X}_{..} + u_{0j} \ + r_{ij}. \end{aligned} \tag{1.6}$$

Then, instead of using $\overline{X}_{.j}$, suppose we have a latent group mean for the jth school. The latent group mean for the jth school $\xi_{.j}$ can replace $\overline{X}_{.j}$, yielding

$$Y_{ij} = \beta_0 + \beta_w X_{ij} + \beta_c \xi_{.j} - \beta_b \xi_{..} + u_{0j} + r_{ij}. \tag{1.7}$$

From equation (1.7), we also assume a simple two-level model for the predictor:

$$\begin{aligned} X_{ij} &= \xi_{.j} + \epsilon_{ij}, \epsilon_{ij} \sim N(0, \sigma_{xx}) \\ \xi_{.j} &= \xi_{..} + \delta_{.j}, \delta_{.j} \sim N(0, \tau_{xx}) \end{aligned} \tag{1.8}$$

The student-level X_{ij} varies around its school mean $\xi_{.j}$. The deviations ϵ_{ij}, follows a normal distribution with mean zero and within-school variance (σ_{xx}). Similarly, the school mean varies around the latent grand mean $(\xi_{..})$. The deviations $(\delta_{.j})$ also follow a normal distribution with a zero mean and between-school variance (τ_{xx}). The bias that is associated with Equation (1.6) instead of using Equation (1.7) can be easily seen when we take expectations of Y_{ij} given X_{ij} and $\overline{X}_{.j}$. The conditional expectation of Equation (1.7) is as follows.

$$E[Y_{ij}|X_{ij}, \overline{X}_{.j}] = \beta_0 + \beta_w X_{ij} + \beta_c E[\xi_{.j}|X_{ij}, \overline{X}_{.j}] - \beta_b E[\xi_{..}|X_{ij}, \overline{X}_{.j}]. \tag{1.9}$$

Using equation (1.8), the expectation of $\xi_{.j}$ given X_{ij} and $\overline{X}_{.j}$ can be written as,

$$E[\xi_{.j}|X_{ij}, \overline{X}_{.j}] = \lambda_j \overline{X}_{.j} + (1 - \lambda_j)\xi_{..}, \tag{1.10}$$

where λ_j=$\tau_{xx}/(\tau_{xx}$+$\sigma_{xx}/n_j)$ is called the "reliability" of $\overline{X}_{.j}$ as an estimate of $\xi_{.j}$ (Raudenbush and Bryk, 2002, chap.3). If we insert Equation (1.10) into Equation (1.9) and rearrange terms, we obtain the following equation:

$$E[Y_{ij}|X_{ij}, \overline{X}_{.j}] = [\beta_0 + \beta_c(1 - \lambda)\xi_{..}] + \beta_w X_{ij} + \beta_c \lambda \overline{X}_{.j} - \beta_b E[\xi_{..}|X_{ij}, \overline{X}_{.j}]. \tag{1.11}$$

When sample size n_j approaches infinity and λ_j is close to 1, bias (denoted here as $\beta_c(1 - \lambda)$) in estimating β_c is close to 0. However, with a limited sample size, λ cannot be 1 and also λ_j varies from school to school as a function of the within-school sample size n_j. In Equation (1.11), $\overline{X}_{.j}$ is an error-contaminated measurement of $\xi_{.j}$ even if X_{ij} is a perfectly reliable measure of an individual construct. Therefore, the sample mean is an unreliable estimate of the observed group mean, and this unreliability will generally lead to bias in regression coefficients.

Using $\overline{X}_{.j}$ instead of $\xi_{.j}$ can be problematic not only where contextual effects are of interest but also where they need to be controlled for. For example, in the case of quasi-experiments or policy studies where intact classes or schools are assigned to different conditions, researchers often want to (or should) control for group mean predictors; otherwise, differences that we see between level-2 units could be due to differences in pre-existing compositional effects rather than the effects of treatment. However, working with $\overline{X}_{.j}$ could result in under adjustments due to regression attenuation. Similarly, cross-level interaction effects could also be underestimated due to attenuation.

1.1.5 Modeling Contextual Effects through Multilevel Latent Variable Modeling

To handle measurement error and sampling error more properly, *multilevel latent variable modeling* has been suggested as an alternative to traditional methods (e.g. Lüdtke et al., 2008; Lüdtke, Marsh, Robitzsch, & Trautwein, 2011; Marsh et al., 2009).

In the present research, multilevel latent variable modeling refers to a class of parametric statistical models that specify linear or nonlinear multilevel relations among a set of continuous latent variables. Different terms such as *multilevel latent structure models* and *multilevel structural equation models* are also used interchangeably. In this statistical modeling framework, observed variables are related to latent variables via the *measurement model*, and the relations among latent variables are defined by multiple levels of the *structural model*.

Lüdtke et al. (2008) proposed a multilevel latent variable modeling framework for contextual analysis. Lüdtke et al. (2008)'s simulation study is noteworthy in that the study examined the relative bias in contextual effect estimates when the traditional HLM model is used under different data conditions. The results showed that the relative percentage bias of contextual effect was less than 10% across varying data conditions when a multilevel latent variable model was used. On the other hand, the relative percentage bias of contextual effect was up to 80% when the traditional HLM model was used.

However, the traditional HLM model can yield less than 10% relative bias under favorable data conditions - that is, when level-1 and level-2 units exceed 30 and 500, respectively, and when there is substantial intra-class correlation (ICC) in the predictor (e.g., 0.3). The study also compared the limited-information approach with full information estimation, and results suggest that the full information estimation is particularly desirable for small numbers of level-1 or level-2 units and a small ICC. However, the manifest variables are limited to only continuous variables in Lüdtke et al. (2008), which is different from the current study. Here, multiple categorical variables are used as manifest variables for both latent predictor and outcome variables. While previous research adopted the EM algorithm with numerical integration for model estimation, the current study adopted an MH-RM algorithm to avoid high dimensional numerical integration and thereby achieve higher efficiency.

Another study using multilevel latent variable modeling for contextual effect analysis was conducted by Marsh et al. (2009). Marsh and colleagues reviewed and compared several contextual modeling options related to BFLPE estimates using an empirical data set in which academic achievement and self-concept were measured by three and four continuous manifest variables, respectively. Among the tested models, a multilevel latent variable model that takes both measurement and sampling error into account yielded the largest BFLPE estimate. The authors described this model as a *doubly latent variable contextual model*. Such a model is theoretically the most desirable choice for researchers, since the model tries to took both measurement and sampling error into account by utilizing information from the manifest variables, rather than using summed or averaged scores of those manifest variables. The study also illustrated how the nonlinear multilevel latent variable modeling approach can provide flexibility in modeling by including random slopes, latent (within-level or cross-level) interactions, and latent quadratic effects. Marsh et al. (2009)'s study used three continuous manifest variables, while the current study considers categorical indicators for all latent variables in the model.

However, nonlinear multilevel latent variable modeling presents significant computational difficulties. Standard approaches such as numerical integration (e.g., adaptive quadrature) or Markov chain Monte Carlo (MCMC, e.g., Gibbs Sampling) based estimation methods have important limitations that make them less practical for routine use, because their computational efficiency drops dramatically when the dimensionality is high. Lüdtke et al. (2011) also reported the occurrence of unstable estimates.

The model has difficulty reaching convergence when sample size is small and the predictors have small intraclass correlations, or when there are substantial amounts of missing observations. Another model specification issue is the assumptions imposed on the distributions of manifest variables. Though it is currently possible to fit a multilevel latent variable model to a real data set that has categorical manifest variables, estimation and model fit diagnosis is more difficult when compared to the cases with continuous manifest variables. Particularly, the underlying contingency table for categorical manifest variables can have many empty cells when the number of categories or items is large and the sample size is small. Therefore, further research is needed to improve estimation of contextual effect in the nonlinear multilevel latent variable modeling framework.

1.2 Research Goals

This study considers a contextual effect not only as a compositional effect that captures the influence of contextual variables on individual level outcomes, but also cross-level interactions that capture the influence of contextual variables on within-group slopes, group-varying compositional effects, and eventually individual-level outcomes.

The main objective of the current study is to develop a more efficient and stable estimation method for contextual effects in the nonlinear multilevel latent variable modeling framework, using Metropolis-Hastings Robbins-Monro algorithm (MH-RM; Cai, 2008, 2010a, 2010b). Computational efficiency and parameter recovery will be assessed in a comparison with EM algorithm using adaptive Gauss-Hermite quadrature for numerical integration (e.g. Mplus; Muthén & Muthén, 2008).

Another objective is to find, through a simulation study, how much measurement error and sampling error can influence contextual effect estimates under different conditions. The results will provide the rationale for using computationally demanding nonlinear multilevel latent variable models. Those conditions cover number of indicators, types of indicators, level-1 and level-2 sample sizes, as well as size of ICCs. Previous research (See, Lüdtke et al., 2011; Marsh et al., 2009) provide guidelines for the varying conditions.

The last objective of the proposed study is to provide an empirical illustration of estimating contextual effects by applying nonlinear multilevel latent variable models to real data that contain more complex measurement structures and unbalanced data. Subsets from Programme for International Student Assessment (PISA; Adams & Wu, 2002) are analyzed to illustrate a contextual effect model and a cross-level interaction model.

1.3 Research Significance

This study is situated in the current streams of research (e.g., Goldstein & Browne, 2004; Goldstein, Bonnet, & Rocher, 2007; Kamata, Bauer, & Miyazaki, 2008) that try to develop a comprehensive, unified model that benefits from both multilevel modeling and latent variable modeling by combining multidimensional IRT and factor analytic measurement modeling with the flexibility of nonlinear structural modeling in a multilevel setting. Considering that one of the most urgent needs in developing a unified model is an efficient estimation method, the current study contributes to nonlinear multilevel latent variable modeling by investigating an alternative estimation algorithm. The principles of MH-RM algorithm and the previous study results (Cai, 2008) suggest that the algorithm can be more efficient than the existing algorithms when a model is associated with a large number of latent variables or random effects.

As computational breakthroughs have contributed to wide applications of statistical models (e.g., EM algorithm), a computational contribution can benefit both methodological researchers and substantive researchers by making nonlinear multilevel latent variable modeling more practically applicable. The current study is expected to make contributions to statistical modeling in educational research by providing discussions on how measurement error and sampling error can affect estimates in multilevel models. The estimation of contextual effects using multilevel latent variable modeling is associated with more precise estimation of group-level latent means (e.g. class- or school-level achievement or teacher characteristics, and environmental characteristics). Obtaining precise group-level latent means is particularly important in estimating teacher or school effects (e.g., value-added models) since many important educational decisions are made based on these results (e.g., budget allocation, school shut-down). Additionally, the multilevel latent variable modeling framework is useful in that this approach takes measurement and sampling error into account properly. This is significant not only when the contextual effects are of interest, but also when they need to be statistically controlled for, as in the case of quasi-experiments or policy evaluation studies.

Furthermore, developments in statistical modeling provide researchers with opportunities to contemplate more refined meaning of contextual effects as compositional effects as well as cross-level interactions. For example, the meaning of compositional effects in the traditional HLM framework is different from those in nonlinear multilevel latent variable modeling when a cross-level interaction is considered.

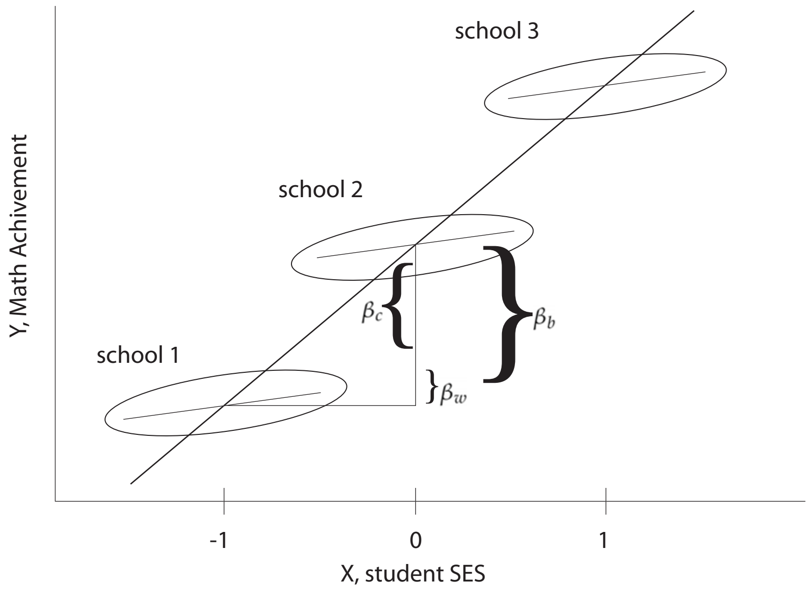


Figure 1.1: Illustration of the compositional effect (β_c) associated with attending school 2 versus school 1

(Raudenbush & Bryk, 2002)

CHAPTER 2

Nonlinear Multilevel Latent Variable Model

This chapter provides some theoretical background on nonlinear multilevel latent variable modeling along with estimation methods. The chapter also describes the observed and complete data likelihoods of contextual models that are necessary to obtain the maximum likelihood estimate (MLE) of parameters using an MH-RM algorithm.

2.1 Development of Nonlinear Multilevel Latent Variable Model

2.1.1 Structural Equation Modeling

Structural equation modeling is rooted in path analysis and factor analysis (Bollen, 1989). Path analysis (Wright, 1918, 1921, 1934, 1960) contributed the path diagrams and the equations that relate correlations or covariances to parameters in current structural equation modeling. Factor analysis (Spearman, 1904) contributed to the conceptual synthesis of latent variable and measurement models in structural equation modeling. With some exceptions, such as the EQS model (Bentler, 1985), factor analysis in general governs the measurement part of latent variable modeling, and path analysis deals with the structural relationship among latent variables. Though the origins of this analytical framework go back to the early 1900's, applications of latent structure models that contain both measurement models and linear structural equations has become prevalent since the 1970's. The breakthroughs by Keesling (1972), Jöreskog (1973), and Wiley (1973) made the practical applications possible. Starting with the LISREL program (Jöreskog & Sörbom, 1974), software packages such as EQS (Bentler, 1985) and Mplus (Muthén & Muthén, 2010) contributed to the increased popularity of structural equation modeling with latent variables.

2.1.2 Multilevel Structural Equation Modeling

Structural equation modeling with latent variables in the multilevel context emerged from the hopes of combining the best of multilevel and latent variable modeling (e.g. McDonald & Goldstein, 1989; McDonald, 1993, 1994; Lee, 1990; Lee & Poon, 1998; Muthén, 1990; Muthén, 1991; Muthén, 1994; Raudenbush & Willms, 1995). McDonald and Goldstein (1989) and McDonald (1993, 1994) focused on estimation in the case of both balanced and unbalanced designs. Muthén (1990), Muthén (1991), and Muthén (1994) proposed a partial maximum likelihood solution in the case of unbalancedness, and Lee and Poon (1998), Raudenbush and Willms (1995), and Liang and Bentler (2004) developed a full maximum likelihood estimator using the EM algorithm. Recent research efforts for multilevel structural equation modeling with latent variables are moving toward unifying and extending generalized linear mixed models, multilevel factor and item response models, and multilevel structural equation models (e.g., Rabe-Hesketh, Skrondal, & Pickles, 2004; Skrondal & Rabe-Hesketh, 2004; Goldstein & Browne, 2004; Goldstein et al., 2007; Kamata et al., 2008). This stream of research is also observed in measurement theory frameworks, e.g., multilevel IRT, multilevel factor analysis framework (Adams, Wilson, & Wu, 1997; Ansari & Jedidi, 2000; Kamata, 2001; Fox & Glas, 2001; Maier, 2001), and explanatory IRT (de Boeck & Wilson, 2004).

2.1.3 Nonlinear Multilevel Structural Equation Modeling

Another important point that stands out in nonlinear multilevel latent structure modeling is *nonlinearity* in two different parts of a contextual model. The measurement model can be nonlinear. Take, for example, de Boeck and Wilson (2004)'s illustrations of the ways in which IRT models can also be considered nonlinear random effects models. In addition to measurement models, nonlinear terms could be directly specified in structural models to accommodate interaction or polynomial effects (Kenny & Judd, 1984).

When modeling contextual effects, the variability in within-group slopes or contextual effects can be explained by a cross-level interaction term, which introduces nonlinearity into the structural model. Bauer and Cai (2009) reported that omitting nonlinearity in multilevel modeling can result in spurious random variation in regression slopes and cross-level interactions. Nonlinear functional forms have been studied from both multilevel (e.g. Cudeck & du Toit, 2003) and structural equation modeling perspectives (e.g. Arminger & Muthén, 1998; Cudeck, Harring, & du Toit, 2009; Lee, Song, & Poon, 2004). However, the models still impose heavy computational burden because of high dimensionality in the latent variable space.

2.2 Contextual Effects in a Nonlinear Multilevel Latent Variable Model

2.2.1 Latent Structure Models

For a contextual effect as a compositional effect that is based on latent variables, we can start with Equation (1.1). Recall that Y_{ij} and X_{ij} denote the outcome and predictor values of student i in school j, respectively. Instead of using Y_{ij} and X_{ij} that are observed variables, we substitute them with latent variables η_{ij} and ξ_{ij} for individual i in group j. Then Equation (1.1) translates into the following:

$$\begin{aligned}
\eta_{ij} &= \beta_{0j} + \beta_{1j}(\xi_{ij} - \xi_{.j}) + r_{ij}, \\
\beta_{0j} &= \gamma_{00} + \gamma_{01}(\xi_{.j} - \xi_{..}) + u_{0j}, \\
\beta_{1j} &= \gamma_{10}, \\
\gamma_{10} &= \beta_{w}, \\
\gamma_{01} &= \beta_{b}, \\
\beta_{c} &= \gamma_{01} - \gamma_{10} \qquad (2.1)
\end{aligned}$$

Similar to Equation (1.1), the random effects r_{ij} and u_{0j} are assumed to be normally distributed with zero means and variances σ^2 and τ_{00}, respectively.

To model varying within-group slopes (β_{1j}), a random effect (u_{1j}) can be added, re-defining β_{1j} as the following:

$$\begin{aligned}
\eta_{ij} &= \beta_{0j} + \beta_{1j}(\xi_{ij} - \xi_{.j}) + r_{ij}, \\
\beta_{0j} &= \gamma_{00} + \gamma_{01}(\xi_{.j} - \xi_{..}) + u_{0j}, \\
\beta_{1j} &= \gamma_{10} + u_{1j}, \\
\gamma_{10} &= \beta_w, \\
\gamma_{01} &= \beta_b, \\
\beta_c &= \gamma_{01} - \gamma_{10} \qquad (2.2)
\end{aligned}$$

In Equation (2.2), u_{1j} also follows a normal distribution with mean zero and variance τ_{11}. The covariance between u_{0j} and u_{1j} is τ_{10}. For both models, the contextual effect β_c is defined as a compositional effect, that is, the difference between γ_{01} and γ_{10}.

However, care must be taken to the contextual effect, particularly when the within-group slopes are random. For example, the contextual effect might also be treated as a random effect that varies across the groups. Therefore, the interpretation of the difference between γ_{01} and γ_{10} in Equation (2.2) is not exactly the same as the interpretation of the compositional effect in Equation (2.1).

Now consider a contextual effect as a cross-level interaction. The grand-mean-centered contextual variable $(\xi_{.j})$ is included in the model as a predictor for β_{1j}. Therefore, β_{1j} is re-defined as follows:

$$\begin{aligned}
\eta_{ij} &= \beta_{0j} + \beta_{1j}(\xi_{ij} - \xi_{.j}) + r_{ij}, \\
\beta_{0j} &= \gamma_{00} + \gamma_{01}(\xi_{.j} - \xi_{..}) + u_{0j}, \\
\beta_{1j} &= \gamma_{10} + \gamma_{11}(\xi_{.j} - \xi_{..}) + u_{1j}, \\
\gamma_{10} &= \beta_w, \\
\gamma_{01} &= \beta_b, \\
\beta_c &= \gamma_{01} - \gamma_{10} \qquad (2.3)
\end{aligned}$$

Again, the compositional effect β_c is considered only when the cross-level interaction of the predictor is controlled for. Even a simple compositional effect can be defined in several different ways depending on what kind of aspects are considered and the research question being answered. For notational simplicity, latent individual deviations from latent group means $(\xi_{ij} - \xi_{.j})$ can be defined as δ_{ij}, and group mean deviations from the latent grand mean $(\xi_{.j} - \xi_{..})$ can be defined as $\delta_{.j}$. Equation (2.3) can be re-written as:

$$\begin{aligned}
\eta_{ij} &= \beta_{0j} + \beta_{1j}\delta_{ij} + r_{ij}, \\
\beta_{0j} &= \gamma_{00} + \gamma_{01}\delta_{.j} + u_{0j}, \\
\beta_{1j} &= \gamma_{10} + \gamma_{11}\delta_{.j} + u_{1j}, \\
\gamma_{10} &= \beta_w, \\
\gamma_{01} &= \beta_b, \\
\beta_c &= \gamma_{01} - \gamma_{10} \qquad (2.4)
\end{aligned}$$

Substituting level-2 effects into the level-1 equation, the reduced form equation is:

$$\eta_{ij} = \gamma_{00} + \gamma_{01}\delta_{.j} + \gamma_{10}\delta_{ij} + \gamma_{11}\delta_{.j}\delta_{ij} + u_{1j}\delta_{ij} + u_{0j} + r_{ij} \qquad (2.5)$$

In Equation (2.5), it is more transparent that the difference between two coefficients γ_{01} and γ_{10} defines a contextual effect as a compositional effect, and γ_{11} captures a contextual effect as a cross-level interaction effect. By fixing γ_{11} at zero or fixing both γ_{11} and u_{1j} at zero, the traditional contextual model with latent variables can be obtained. Thus the models in Equations (2.1) and (2.2) are nested within the model in Equation (2.5). The three models represented by Equations (2.1), (2.2), and (2.3) are illustrated using path diagrams through Figures 2.1, 2.2, and 2.3.

2.2.2 Measurement Models

The measurement models define the relationship between observed variables and latent variables. The measurement models are developed as IRT models. For brevity, only the measurement models of level-1 latent predictor variable ξ_{ij} will be described in this section, since the measurement models for other variables such as the latent outcome η_{ij} follow the same principles. Let the vector of responses form the ith respondent in the jth group to the set of L observed variables for latent variable ξ_{ij} be $\mathbf{x}_{ij} = (x_{1ij},...,x_{lij},...,x_{Lij})'$. We assume the conditional independence of observed variables given latent trait ξ_{ij} (Lord & Novick, 1968). The likelihood of observing $\mathbf{x}_{ij}$ given ξ_{ij} is:

$$f_\theta(\mathbf{x}_{ij}|\xi_{ij}) = \prod_{l=1}^{L} f_\theta(x_{ijl}|\xi_{ij}), \tag{2.6}$$

where θ contains the free item parameters. The two models considered in this study are for dichotomously scored items and graded response items, but other IRT models may be used.

2.2.2.1 Dichotomous Response

This model can be considered a generalized 2-parameter logistic model (2-PL) as well as a special case of a graded response model with two categories, (to be described in the next section). The conditional probability for x_{ijl} = 1 is

$$P_\theta(x_{ijl} = 1|\xi_{ij}) = \frac{1}{1 + \exp[-(b_l + a_l\xi_{ij})]}, \tag{2.7}$$

Here, b_l and a_l denote intercept (difficulty parameter) and slope (discrimination parameter), respectively. Let χ_k is an indicator function. χ_k is 1 if $x_{ijl} = k$, or 0 otherwise. In the case of dichotomous response, k is 1; therefore, χ_1 is 1 if $x_{ijl} = 1$, or 0 otherwise. Finally, the conditional density for x_{ijl} is that of a Bernoulli random variable:

$$f_\theta(x_{ijl}|\xi_{ij}) = \prod_{k=0}^{1} P_\theta(x_{ijl} = k|\xi_{ij})^{\chi_k(x_{ijl})}, \tag{2.8}$$

θ shows that the item properties belong to the list of free parameters.

2.2.2.2 Graded Responses

When manifest variables are graded response variables with multiple categories, Samejima (1969)'s model can be utilized. Let $x_{1ij} \in \{0, 1, 2, ..., K_l - 1\}$ be an element of ith individual's response in jth group to lth item that has K_l ordered categories. Then the logistic conditional cumulative response probability for each category are listed as follows:

$$\begin{aligned}
P_\theta(x_{ijl} \geq 0|\xi_{ij}) &= 1, \\
P_\theta(x_{ijl} \geq 1|\xi_{ij}) &= \frac{1}{1 + \exp[-(b_{1,l} + a_l\xi_{ij})]}, \\
P_\theta(x_{ijl} \geq 2|\xi_{ij}) &= \frac{1}{1 + \exp[-(b_{2,l} + a_l\xi_{ij})]}, \\
&\vdots \\
P_\theta(x_{ijl} \geq K_l - 1|\xi_{ij}) &= \frac{1}{1 + \exp[-(b_{K_l-1,l} + a_l\xi_{ij})]},
\end{aligned} \tag{2.9}$$

The category response probability is defined as the difference between two adjacent cumulative probabilities:

$$P_\theta(x_{ijl} = k|\xi_{ij}) = P_\theta(x_{ijl} \geq k|\xi_{ij}) - P_\theta(x_{ijl} \geq k+1|\xi_{ij}), \tag{2.10}$$

where $P_\theta(x_{ijl} \geq k|\xi_{ij})$ is zero. Again, χ_k is an indicator function in which χ_k is 1 if $x_{ijl} = k$, or 0 otherwise. The conditional density for x_{ijl} follows a multinomial with trial size 1 in K_l categories:

$$f_\theta(x_{ijl}|\xi_{ij}) = \prod_{k=0}^{K_l-1} P_\theta(x_{ijl} = k|\xi_{ij})^{\chi_k(x_{ijl})}. \tag{2.11}$$

2.2.3 Observed and Complete Data Likelihoods

As ξ_{ij} is measured by $\mathbf{x}_{ij}$, η_{ij} is measured by $\mathbf{y}_{ij}$, the conditional density of $\mathbf{y}_{ij}$ is written as:

$$f_\theta(\mathbf{y}_{ij}|\eta_{ij}) = f_\theta(\mathbf{y}_{ij}|\xi_{ij}, \xi_{.j}, \boldsymbol{\beta}_j, r_{ij}), \tag{2.12}$$

If we integrate r_{ij} out of Equation (2.12),

$$\int f_\theta(\mathbf{y}_{ij}|\xi_{ij}, \xi_{.j}, \boldsymbol{\beta}_j) f_\theta(r_{ij}) d(r_{ij}) = f_\theta(\mathbf{y}_{ij}|\xi_{ij}, \xi_{.j}, \boldsymbol{\beta}_j), \tag{2.13}$$

where $f_\theta(r_{ij})$ is the density of a normal distribution $N(0, \sigma^2)$. For identification purpose, σ^2 is fixed at 1 in this study, which makes $f_\theta(r_{ij})$ the density of a standard normal random variable. Integrating out ξ_{ij} yields

$$\begin{aligned} & f_\theta(\mathbf{y}_{ij}, \mathbf{x}_{ij}|\xi_{.j}, \boldsymbol{\beta}_j) \\ = & \int f_\theta(\mathbf{x}_{ij}|\xi_{ij}) f_\theta(\mathbf{y}_{ij}|\xi_{ij}, \xi_{.j}, \boldsymbol{\beta}_j) f(\xi_{ij}) d(\xi_{ij}) \end{aligned} \tag{2.14}$$

When J and I_j stand for the number of groups and number of individuals in group j, the conditional joint density of $\mathbf{y}_{.j}$ and $\mathbf{x}_{.j}$ for group j is the multiplication of the conditional joint densities for $\mathbf{y}_{ij}$ and $\mathbf{x}_{ij}$ in the same group as can be seen in the following equation:

$$f_\theta(\mathbf{y}_{.j}, \mathbf{x}_{.j}|\xi_{.j}, \boldsymbol{\beta}_j) = \prod_{i=1}^{I_j} f_\theta(\mathbf{y}_{ij}, \mathbf{x}_{ij}|\xi_{.j}, \boldsymbol{\beta}_j) \tag{2.15}$$

Integrating out level-2 latent variable and random coefficients $\xi_{.j}$ and $\boldsymbol{\beta}_j$ yields

$$f_\theta(\mathbf{y}_{.j}, \mathbf{x}_{.j}) = \int \prod_{i=1}^{I_j} f_\theta(\mathbf{y}_{ij}, \mathbf{x}_{ij}|\xi_{.j}, \boldsymbol{\beta}_j) f(\xi_{.j}) f(\boldsymbol{\beta}_j) d(\xi_{.j}) d(\boldsymbol{\beta}_j) \tag{2.16}$$

In this manner, one can integrate all latent variables and random coefficients out of the model to get a marginal distribution from which the parameters can be estimated. Treating η_{ij}, ξ_{ij}, $\xi_{.j}$, $\boldsymbol{\beta}_j$ and r_{ij} as missing data, the complete data likelihood, when J and

I_j stand for the number of groups and number of individuals in group j, is:

$$\prod_{j=1}^{J}\Big[\prod_{i=1}^{I_j} f_\theta(\mathbf{y}_{ij}|\xi_{ij},\xi_{.j},\boldsymbol{\beta}_j,r_{ij})f_\theta(\mathbf{x}_{ij}|\xi_{ij})f_\theta(\xi_{ij})f_\theta(r_{ij})\Big]\times f_\theta(\boldsymbol{\beta}_j)f_\theta(\xi_{.j}) \tag{2.17}$$

where $f_\theta(\mathbf{x}_{ij}|\xi_{ij})=\prod_{l=1}^{L_x} f_\theta(x_{ijl}|\xi_{ij})$ and $f_\theta(\mathbf{y}_{ij}|\boldsymbol{\xi}_{ij},\boldsymbol{\xi}_{.j},\boldsymbol{\beta}_j)=\prod_{l=1}^{L_y} f_\theta(y_{ijl}|\boldsymbol{\xi}_{ij},\boldsymbol{\xi}_{.j},\boldsymbol{\beta}_j)$. L_x and L_y are the number of manifest variables for ξ_{ij} and η_{ij}, respectively.

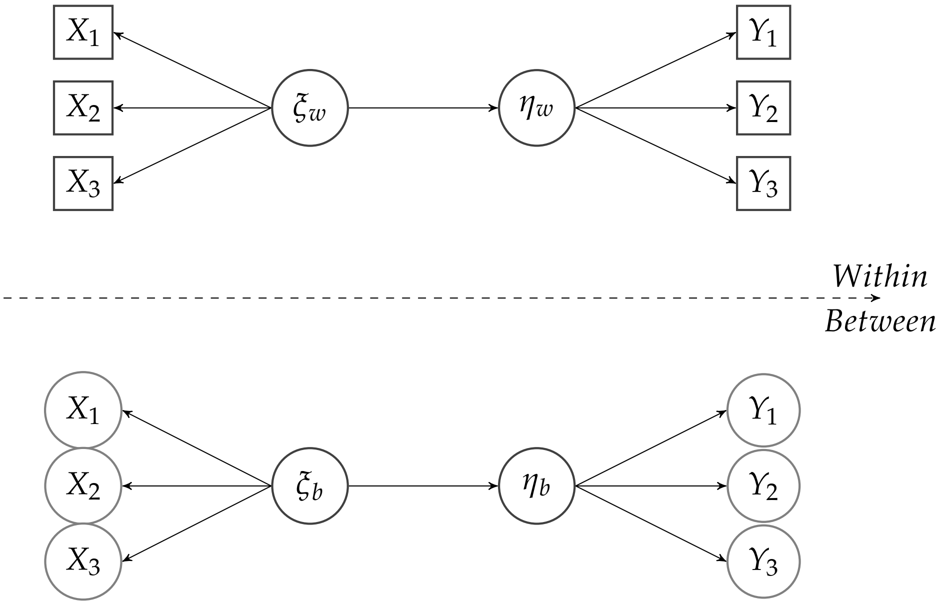


Figure 2.1: Conceptual path diagram showing the compositional effect model in Equation (2.1).

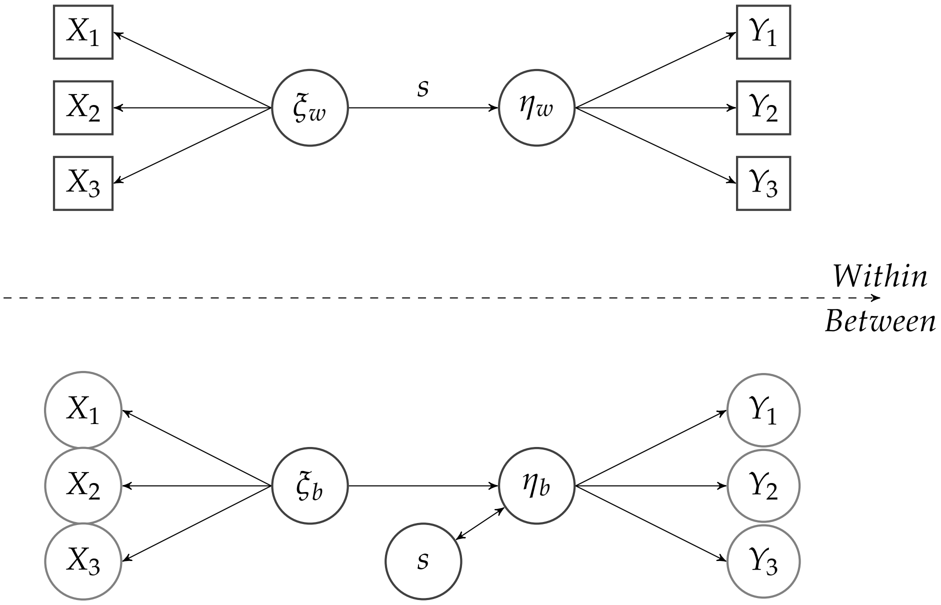


Figure 2.2: Conceptual path diagram showing the compositional effect model with random slopes in Equation (2.2).

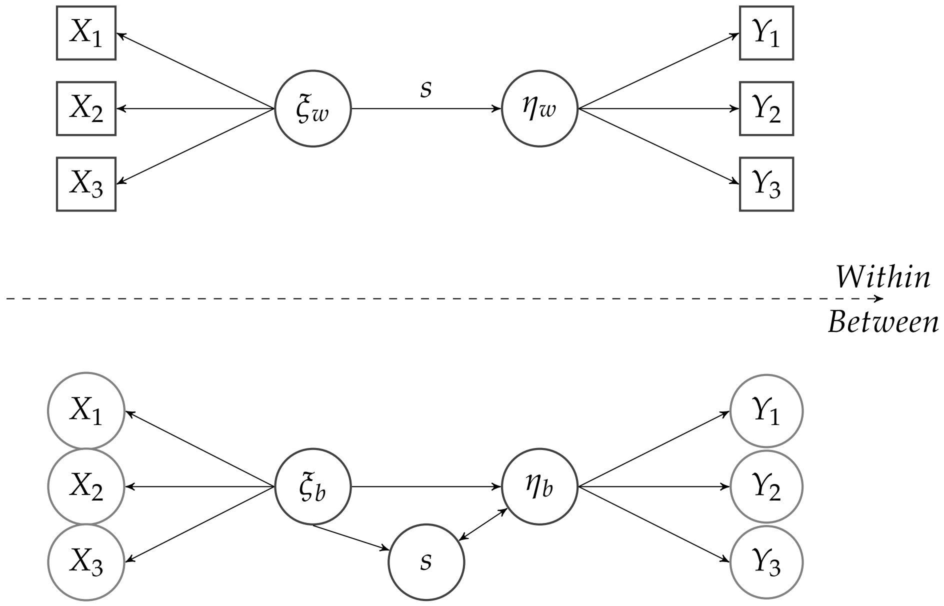


Figure 2.3: Conceptual path diagram showing the compositional effect model with random slopes and a cross-level interaction in Equation (2.3).

CHAPTER 3

A Metropolis-Hastings Robbins-Monro Algorithm

Considering the missing data formulation, where the observed data are $\mathbf{Y}_o$ and the missing data are $\mathbf{Y}_m$, the observed data likelihood can be written as $L(\theta|\mathbf{Y}_o)$ and the complete data likelihood function is $L(\theta|\mathbf{Y})$ where $\mathbf{Y} = (\mathbf{Y}_o,\mathbf{Y}_m)$. While maximizing $L(\theta|\mathbf{Y}_o)$ involves high-dimensional integrals, the complete data likelihood $L(\theta|\mathbf{Y})$ involves a series of products of likelihoods that are fairly simple to maximize. Therefore, having plausible values of random effects and latent variables makes the estimation problem simpler. This also allows straightforward optimization of the the complete data likelihood with respect to θ. However, proper imputation requires the distribution of the missing data to be conditional on the observed data. As the model is nonlinear, analytical derivation of the distribution of missing data conditional on the observed data is difficult. Nevertheless, a property of the posterior of the missing data enables us to have appropriate imputation. That is, the posterior of missing data, given observed data and a provisional θ, is proportional to the complete data likelihood. To utilize this property, Metropolis-Hastings sampler (MH; Hastings, 1970; Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953) is adopted to produce the imputations from a Markov chain with the missing data posterior as the target. Then, the random imputations are combined into Stochastic Approximation using the Robbins-Monro algorithm (RM; Robbins & Monro, 1951).

3.1 The EM Algorithm and MH-RM

Cai (2008) described the MH-RM algorithm as an extension of the Stochastic Approximation EM algorithm (SAEM; Celeux & Diebolt, 1991; Celeux, Chauveau, & Diebolt, 1995; Delyon, Lavielle, & Moulines, 1999). Accordingly, it is helpful to review the conventional EM algorithm for incomplete data before proceeding to the MH-RM algorithm.

The EM algorithm is composed of two steps (Dempster, Laird, & Rubin, 1977). The E-step computes the conditional expectation of complete data log-likelihood using the current estimates for the parameters, and the M-step maximizes the conditioned expected log-likelihood found in the E-step. By alternating the two steps until convergence, maximum likelihood or maximum a posteriori (MAP) estimates of parameters are obtained.

Using the missing data notation used in the previous chapter, the complete data can be written as $\mathbf{Y} = (\mathbf{Y}_o, \mathbf{Y}_m)$. The complete data and observed data log-likelihood can be expressed as $l(\boldsymbol{\theta}|\mathbf{Y})$ and $l(\boldsymbol{\theta}|\mathbf{Y}_o)$, respectively.

When the current estimate is denoted as $\boldsymbol{\theta}^*$, the expected complete-data log-likelihood

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^*) = \int (l(\boldsymbol{\theta}|\mathbf{Y}))F_{\boldsymbol{\theta}^*}(d\mathbf{Y}_m|\mathbf{Y}_o) \tag{3.1}$$

is computed in E(xpectation)-step where $F_{\boldsymbol{\theta}}(\mathbf{Y}_m|\mathbf{Y}_o)$ denotes the posterior predictive distribution of missing data. Then M(aximization)-step computes new parameters that maximize $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)$.

Fisher (1925) proved that the conditional expectation of $\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}|\mathbf{Y})$, on the right had side of the Equation 3.2, is the same as the gradient of the observed data log-likelihood, on the left hand side of Equation 3.2.

$$\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}|\mathbf{Y}_o) = \int_{\mathcal{E}} \nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}|\mathbf{Y}) F_{\boldsymbol{\theta}}(d\mathbf{Y}_m|\mathbf{Y}_o), \tag{3.2}$$

where $F_{\theta}(\mathbf{Y}_m|\mathbf{Y}_o)$ is the posterior predictive distribution of missing data, and $\mathcal{E}$ is some sample space. Equation (3.2) is known as Fisher's Identity. Cai (2008) pointed out that

Fisher's Identity is a strong motivation of the MH-RM algorithm in that one can obtain the gradient of the observed data log-likelihood from the conditional expectation of the complete data gradient $\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}|\mathbf{Y})$. The solution that makes the right-hand side of Equation (3.2) zero is the same solution that makes $\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}|\mathbf{Y}_o)$ zero.

Taking the expectation of $\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}|\mathbf{Y})$ with respect to $F_{\boldsymbol{\theta}}(\mathbf{Y}_m|\mathbf{Y}_o)$ is now critical. This can be accomplished by imputing missing data from its posterior predictive distribution. As the posterior distribution depends on unknown $\boldsymbol{\theta}$, the solution needs to be obtained iteratively.

3.2 The RM Algorithm and MH-RM

Robbins and Monro (1951)'s algorithm is a root-finding algorithm under observational noise functions and MH-RM can be conceived of as a generalized RM algorithm (Cai, 2008) for multiple parameters. Let $\boldsymbol{\theta}$ be a variable and $g(\cdot)$ be a continuously differentiable function. Newton's procedure yields the following equation to find the root of a function $g(\boldsymbol{\theta})$:

$$\boldsymbol{\theta}^{k+1} = \boldsymbol{\theta}^{k} + [-\nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta}^{k})]^{-1} g(\boldsymbol{\theta}^{k}). \tag{3.3}$$

The procedure starts with $\boldsymbol{\theta}^{k}$, which is a starting value when $k = 0$, and then iteratively updates $\boldsymbol{\theta}^{k}$. When the function is unknown or not differentiable, the following RM recursive filter can be used analogously to Newton's procedure:

$$\boldsymbol{\theta}^{k+1} = \boldsymbol{\theta}^{k} + \gamma_k R_{k+1}, \tag{3.4}$$

where $R_{k+1} = g(\boldsymbol{\theta}^{k})+\zeta_{k+1}$ and γ_k is a sequence of *gain constants*. Here, ζ_{k+1} is a random variable with mean zero. Because $g(\boldsymbol{\theta}^{k})$ is unknown, R_{k+1} becomes the estimate of $g(\boldsymbol{\theta}_k)$. On the other hand, the sequence of gain constants γ_k should satisfy the following conditions:

$$\gamma_k \in (0, 1], \Sigma_{k=1}^{\infty} \gamma_k = \infty, \Sigma_{k=1}^{\infty} \gamma_k^2 < \infty. \tag{3.5}$$

These conditions make the gain constants decrease *slowly* to zero. An interesting part of this approach is that R_{k+1} does not need to be highly accurate since it only provides the right direction for the next move. The role of decaying gain constants is to eliminate the effect of the noise, enabling $\boldsymbol{\theta}^k$ to converge to the root point-wise. The MH-RM algorithm is a generalized form of the RM Algorithm for multiple parameters (Cai, 2008). The noise is introduced by stochastic data augmentation. Recall that we need to take the expectation of $\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}|\mathbf{Y})$ with respect to $F_{\boldsymbol{\theta}}(\mathbf{Y}_m|\mathbf{Y}_o)$. A Markov chain can be constructed to draw plausible missing values from the posterior predictive distribution to obtain complete data $\mathbf{Y}$. Let $\boldsymbol{\theta}^k$ be the estimate at the end of iteration k.

The $(k+1)$th iteration of the MH-RM algorithm consists of 3 steps: Stochastic Imputation, Stochastic Approximation, and Robbins-Monro Update.

1. Stochastic Imputation

Draw m_k sets of missing data, which are the random effects and latent variables, from a Markov chain that has the distribution of missing data conditional on observed data as the target. Then, m_k sets of complete data are as follows:

$$\left\{\mathbf{Y}_j^{k+1}; j = 1, ..., m_k\right\} \tag{3.6}$$

2. Stochastic Approximation

Using Fishier's Identity, a Monte Carlo approximation to $\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}^k|\mathbf{Y}_o)$ can be computed as the sample average of complete data gradients. We also compute a recursive approximation of the conditional expectation of the information matrix of the complete data log-likelihood. For simplicity, let $\mathbf{s}(\boldsymbol{\theta}|\mathbf{Y})$ stand for $\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}|\mathbf{Y})$, and the sample average of complete data gradients can be written as:

$$\tilde{\mathbf{s}}_{k+1} = \frac{1}{m_k} \sum_{j=1}^{m_k} \mathbf{s}(\boldsymbol{\theta}^k|\mathbf{Y}_j^{k+1}), \tag{3.7}$$

and $\mathbf{\Gamma}_{k+1}$ is

$$\mathbf{\Gamma}_{k+1} = \mathbf{\Gamma}_k + \gamma_k \left[\frac{1}{m_k} \sum_{j=1}^{m_k} \mathbf{H}(\boldsymbol{\theta}^k | \mathbf{Y}_j^{k+1}) - \mathbf{\Gamma}_k \right], \tag{3.8}$$

where $\mathbf{H}(\boldsymbol{\theta}|\mathbf{Y})$ is the complete data information matrix, which is -1 times the second derivative matrix of the complete data log-likelihood.

3. Robbins-Monro Update

Now new parameters are estimated through the following update:

$$\boldsymbol{\theta}^{k+1} = \boldsymbol{\theta}^k + \gamma_k (\mathbf{\Gamma}_{k+1}^{-1} \tilde{\mathbf{s}}_{k+1}) \tag{3.9}$$

The iterations can be stopped upon convergence wheen the changes in parameter estimates are sufficiently small. Cai (2008) verified that the asymptotic behaviors of MH-RM in time and it converges to MLE. More detailed information about the relationship between MH-RM and other existing algorithms are described in Cai (2008).

3.3 Approximation to the Observed Information Matrix

One of the benefits of using the MH-RM algorithm is that the observed data information matrix can be approximated as a byproduct of the iterations. The inverse of the observed data information matrix becomes the large-sample covariance matrix of parameter estimates. The square root of the diagonal elements are the standard errors. Utilizing Fishier's Identity, the score vector is approximated recursively at kth iteration,

$$\hat{\mathbf{s}}_{k+1} = \hat{\mathbf{s}}_{k-1} + \gamma_k \{ \tilde{\mathbf{s}}_{k+1} - \hat{\mathbf{s}}_k \}, \tag{3.10}$$

where $\tilde{\mathbf{s}}_k$ is defined as Equation (3.7) and γ_k is a sequence of gain constants. A Monte Carlo estimate of the conditional expectation is defined as follows:

$$\tilde{\mathbf{G}}_k = \frac{1}{m_k} \sum_{j=1}^{m_k} \left[\mathbf{H}(\boldsymbol{\theta}^k | \mathbf{Y}_j^{k+1}) - \mathbf{s}(\boldsymbol{\theta}^k | \mathbf{Y}_j^{k+1}) [\mathbf{s}(\boldsymbol{\theta}^k | \mathbf{Y}_j^{k+1})]' \right]. \tag{3.11}$$

For recursive SA, a better estimate is defined as the next equation since the $\hat{\mathbf{s}}_k$ is too noisy.

$$\hat{\mathbf{G}}_{k+1} = \hat{\mathbf{G}}_k + \gamma_k\{\tilde{\mathbf{G}}_{k+1} - \hat{\mathbf{G}}_k\}. \tag{3.12}$$

Finally, the observed information matrix is approximated as

$$I_{k+1} = \hat{\mathbf{G}}_{k+1} + \hat{\mathbf{s}}_{k+1}\hat{\mathbf{s}}'_{k+1}, \tag{3.13}$$

Another practical option for approximating the observed information matrix is a direct application of Louis's (1982) approach, in which the score vector and the conditional expectation are approximated directly after converge. The straightforward differentiation of Equation 3.1 yields (with slight changes in notations here from Louis's (1982) formula),

$$\begin{aligned} I_{Y_o} = {} & E_{\boldsymbol{\theta}}\Big\{\mathbf{H}(\mathbf{Y},\boldsymbol{\theta})|\mathbf{Y}\Big\} \\ & - E_{\boldsymbol{\theta}}\Big\{\mathbf{s}(\mathbf{Y},\boldsymbol{\theta})\mathbf{s}^T(\mathbf{Y},\boldsymbol{\theta})|\mathbf{Y}\Big\} \\ & + E_{\boldsymbol{\theta}}\Big\{\mathbf{s}(\mathbf{Y},\boldsymbol{\theta})|\mathbf{Y}\Big\}E_{\boldsymbol{\theta}}\Big\{\mathbf{s}^T(\mathbf{Y},\boldsymbol{\theta})|\mathbf{Y}\Big\}. \end{aligned} \tag{3.14}$$

$E_{\boldsymbol{\theta}}\Big\{\mathbf{s}(\mathbf{Y},\boldsymbol{\theta})|\mathbf{Y} \in \mathbf{R}\Big\}$ is 0 when $\boldsymbol{\theta}$ is evaluated at MLE $\hat{\boldsymbol{\theta}}$. When v denotes the number of samples that are used to approximate the covariance matrix, and $\mathbf{Y}_i$ is an imputation from $F_\theta(\mathbf{Y}_m|\mathbf{Y}_o)$, the first two terms in Equation 3.14 are calculated using Equations 3.15 and 3.16, respectively.

$$E_{\boldsymbol{\theta}}\Big\{\mathbf{H}(\mathbf{Y},\boldsymbol{\theta})|\mathbf{Y}\Big\} \approx \frac{1}{v}\sum_{i=1}^{v}\mathbf{H}(\hat{\boldsymbol{\theta}},\mathbf{Y}_i|\mathbf{Y}_i), \tag{3.15}$$

$$E_{\boldsymbol{\theta}}\Big\{\mathbf{s}(\mathbf{Y},\boldsymbol{\theta})\mathbf{s}^T(\mathbf{Y},\boldsymbol{\theta})|\mathbf{Y}\Big\} \approx \frac{1}{v}\sum_{i=1}^{v}[\mathbf{s}(\hat{\boldsymbol{\theta}},\mathbf{Y}_i)\mathbf{s}^T(\hat{\boldsymbol{\theta}},\mathbf{Y}_i)|\mathbf{Y}_i]. \tag{3.16}$$

In this study, the first method is called *recursively approximated standard errors* and the latter is called *post-convergence approximated standard errors*. More precisely, these methods approximate the observed data information matrix that yields the standard error estimates. Both methods were adopted for this study to examine the quality of the estimates and practicability.

CHAPTER 4

Implementation of MH-RM for Contextual Models

This chapter describes how an MH-RM algorithm is implemented to obtain maximum likelihood estimates for a contextual effect model and a cross-level interaction model in the multilevel latent variable modeling framework. The first section explains how an MH sampler can be constructed, and the second section reports the complete data models and their first and second derivatives that are used to update parameter estimates. The final section provides details related to acceleration and convergence of the algorithm.

4.1 A Metropolis-Hastings Sampler

The first step of an MH-RM algorithm for multilevel latent variable modeling is the stochastic imputation of latent variables and random effects. The imputation process is composed of 1) generating the candidate values for random effects with a random walk sampler, 2) evaluating acceptance probabilities, and 3) accepting or rejecting the candidates. This process ultimately aims at sampling from the distribution of missing data given observed data $F_{\boldsymbol{\theta}}(\mathbf{Y}_m|\mathbf{Y}_o)$ in Equation (3.2), which is proportional to the complete data likelihood $L(\boldsymbol{\theta}|\mathbf{Y})$. Recall the notation in Equations (2.1), (2.2), and (2.3), where ξ_{ij}, $\xi_{.j}$, $\xi_{..}$ and η_{ij}, are latent variables and r_{ij}, β_{0j}, and β_{1j} are random variables. On the other hand, γ_{00}, γ_{01}, γ_{10}, and γ_{11} are fixed parameters in the structural model, and there are also item parameters in the measurement model. Except for latent variables and random effects, all other parameters are considered fixed in the population and can be denoted as $\boldsymbol{\theta}$. In applications of latent variable modeling in which a latent variable is measured by multiple manifest variables, either *factor standardization* or *anchoring one of the factor* *loadings* should be made for identification purpose. For this study, factor standardization is chosen to estimate all factor loadings. Accordingly, γ_{00} is fixed at zero. Among latent variables, the mean of $\xi_{..}$ can similarly be fixed at zero for identification. Following standard practice in IRT, ξ_{ij} is defined scale with mean zero and variance 1. Furthermore, η_{ij} is a combination of other random variables and can be calculated once ξ_{ij}, $\xi_{.j}$, β_{0j}, β_{1j}, and r_{ij} are known. If we omit η_{ij} and $\xi_{..}$ for now, the remaining random variables of interest are the latent predictor ξ_{ij}, the group level latent predictor $\xi_{.j}$, residuals at level-1 r_{ij}, the within-group intercept β_{0j}, and within-group slope β_{1j}. These variables are viewed as missing data. Therefore, the set of missing data corresponds to $\mathbf{Y}_m$ in Equation (3.2).

Considering the multilevel structure of the proposed model, level-1 latent variables are independent conditional on level-2 latent variables, and level-2 random variables are also independent conditional on level-1 latent variables when there is no cross-level interaction. For further illustration, the vector of level-1 latent variables (ξ_{ij}, r_{ij}) is called $\mathbf{Y}_{m,ij}$, and the vector of level-2 latent variables ($\xi_{.j}$, β_{0j}, β_{1j}) is called $\mathbf{Y}_{m,.j}$. The latent variables $\mathbf{Y}_{m,ij}$ and $\mathbf{Y}_{m,.j}$ are treated as missing data. The MCMC imputation procedure can be constructed using Gibbs sampling. Let $\mathbf{Y}^l_{m,ij}$ be the value of $\mathbf{Y}_{m,ij}$ in the lth iteration of a Gibbs sampler with the following steps:

$$\begin{aligned}
\text{Draw}\mathbf{Y}^l_{m,1j} &\sim f_{\boldsymbol{\theta}}(\mathbf{Y}_{m,1j}|\mathbf{Y}^{l-1}_{m,2j}, ..., \mathbf{Y}^{l-1}_{m,Ij}, \mathbf{Y}_{m,.j}, \mathbf{Y}_o) \\
\text{Draw}\mathbf{Y}^l_{m,2j} &\sim f_{\boldsymbol{\theta}}(\mathbf{Y}_{m,2j}|\mathbf{Y}^{l}_{m,1j}, \mathbf{Y}^{l-1}_{m,3j}, ..., \mathbf{Y}^{l-1}_{m,Ij}, \mathbf{Y}_{m,.j}, \mathbf{Y}_o) \\
&\vdots \\
\text{Draw}\mathbf{Y}^l_{m,ij} &\sim f_{\boldsymbol{\theta}}(\mathbf{Y}_{m,ij}|\mathbf{Y}^{l}_{m,1j}, ..., \mathbf{Y}^{l}_{m,i-1j}, \mathbf{Y}^{l-1}_{m,i+1j}, ..., \mathbf{Y}^{l-1}_{m,Ij}, \mathbf{Y}_{m,.j}, \mathbf{Y}_o) \\
&\vdots \\
\text{Draw}\mathbf{Y}^l_{m,Ij} &\sim f_{\boldsymbol{\theta}}(\mathbf{Y}_{m,Ij}|\mathbf{Y}^{l}_{m,1j}, ..., \mathbf{Y}^{l}_{m,I-1j}, \mathbf{Y}_{m,.j}, \mathbf{Y}_o) \qquad (4.1)
\end{aligned}$$

Each of the full conditionals are still difficult to sample directly. This suggests coupling the Gibbs sampler with the MH algorithm. Let

$$\begin{aligned}&\alpha(\mathbf{Y}_{m,ij},\mathbf{Y}^*_{m,ij}|\theta,\mathbf{Y}_{o,ij},\mathbf{Y}_{m,.j})\\ &=\min\left\{\frac{f_\theta(\mathbf{Y}_{o,ij}|\mathbf{Y}^*_{m,ij})h_{1j}(\mathbf{Y}^*_{m,ij}|\boldsymbol{\mu}_{1j},\boldsymbol{\Sigma}_{1j})h_2(\mathbf{Y}_{m,.j}|\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)q(\mathbf{Y}^*_{m,ij},\mathbf{Y}_{m,ij})}{f_\theta(\mathbf{Y}_{o,ij}|\mathbf{Y}_{m,ij})h_{1j}(\mathbf{Y}_{m,ij}|\boldsymbol{\mu}_{1j},\boldsymbol{\Sigma}_{1j})h_2(\mathbf{Y}_{m,.j}|\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)q(\mathbf{Y}_{m,ij},\mathbf{Y}^*_{m,ij})},1\right\}\end{aligned}\tag{4.2}$$

be the acceptance probability of moving from state $\mathbf{Y}_{m,ij}$ to $\mathbf{Y}^*_{m,ij}$ given parameters $\boldsymbol{\theta}$, observed data $\mathbf{Y}_{o,ij}$, and $\mathbf{Y}_{m,.j}$, where $q(\mathbf{Y}_{m,ij},\mathbf{Y}^*_{m,ij})$ is a transition density. When a simple random walk chain is used, a candidate is $\mathbf{Y}^*_{m,ij}=\mathbf{Y}_{m,ij}+\mathbf{e}_{ij}$, where $\mathbf{e}_{ij}$ follows a scaled multivariate standard normal distribution in p dimensions, where p is the number of latent variables. For example, $p=2$ for the increment to $\mathbf{Y}_{m,ij}=(\xi_{ij},r_{ij})'$, and a set of $\mathbf{e}_{ij}$ is drawn from a scaled standard bivariate normal distribution $N_2(0,w^2I_2)$. The w value can be changed to tune the acceptance ratio of the MH chain and generally needs to be smaller than 1 for high-dimensional problems (Cai, 2008). Due to the symmetry of the increment density, $q(\mathbf{Y}_{m,ij},\mathbf{Y}^*_{m,ij})=q(\mathbf{Y}^*_{m,ij},\mathbf{Y}_{m,ij})$, Equation (4.2), yielding the reduced form as follows:

$$\begin{aligned}&\alpha(\mathbf{Y}_{m,ij},\mathbf{Y}^*_{m,ij}|\theta,\mathbf{Y}_{o,ij},\mathbf{Y}_{m,.j})\\ &=\min\left\{\frac{f_\theta(\mathbf{Y}_{o,ij}|\mathbf{Y}^*_{m,ij})h_{1j}(\mathbf{Y}^*_{m,ij}|\boldsymbol{\mu}_{1j},\boldsymbol{\Sigma}_{1j})h_2(\mathbf{Y}_{m,.j}|\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)}{f_\theta(\mathbf{Y}_{o,ij}|\mathbf{Y}_{m,ij})h_{1j}(\mathbf{Y}_{m,ij}|\boldsymbol{\mu}_{1j},\boldsymbol{\Sigma}_{1j})h_2(\mathbf{Y}_{m,.j}|\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)},1\right\}\end{aligned}\tag{4.3}$$

As it can be seen in Equation (4.3), the density function related to level-2 missing data $h_2(\mathbf{Y}_{m,.j}|\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)$ is the same for the current draws and candidate draws. Therefore, Equation (4.3) can be further reduced as:

$$\alpha(\mathbf{Y}_{m,ij},\mathbf{Y}^*_{m,ij}|\theta,\mathbf{Y}_{o,ij},\mathbf{Y}_{m,.j})=\min\left\{\frac{f_\theta(\mathbf{Y}_{o,ij}|\mathbf{Y}^*_{m,ij},\mathbf{Y}_{m,.j})h_{1j}(\mathbf{Y}^*_{m,ij}|\boldsymbol{\mu}_{1j},\boldsymbol{\Sigma}_{1j})}{f_\theta(\mathbf{Y}_{o,ij}|\mathbf{Y}_{m,ij})h_{1j}(\mathbf{Y}_{m,ij}|\boldsymbol{\mu}_{1j},\boldsymbol{\Sigma}_{1j})},1\right\}\tag{4.4}$$

In short, conditional on $\mathbf{Y}_{m,.j}$, a candidate is $\mathbf{Y}^*_{m,ij}=\mathbf{Y}_{m,ij}+\mathbf{e}_{ij}$. The acceptance probabilities are calculated by Equation (4.4) and the candidates are accepted or rejected based on the evaluation.

Similarly, let $\mathbf{Y}^l_{m,.j}$ be the value of $\mathbf{Y}_{m,.j}$ in the lth iteration of a Gibbs sampler with the following steps:

$$
\begin{array}{rcl}
\text{Draw}\mathbf{Y}^{l}_{m,.1} & \sim & f_{\boldsymbol{\theta}}(\mathbf{Y}_{m,.1}|\mathbf{Y}^{l-1}_{m,.2}, ..., \mathbf{Y}^{l-1}_{m,.J}, \{\mathbf{Y}_{m,ij}\}^{I_1}_{i=1}, \mathbf{Y}_{o}) \\
\text{Draw}\mathbf{Y}^{l}_{m,.2} & \sim & f_{\boldsymbol{\theta}}(\mathbf{Y}_{m,.2}|\mathbf{Y}^{l}_{m,.1}, \mathbf{Y}^{l-1}_{m,.3}, ..., \mathbf{Y}^{l-1}_{m,.J}, \{\mathbf{Y}_{m,ij}\}^{I_2}_{i=1}, \mathbf{Y}_{o}) \\
\vdots & & \\
\text{Draw}\mathbf{Y}^{l}_{m,.j} & \sim & f_{\boldsymbol{\theta}}(\mathbf{Y}_{m,.j}|\mathbf{Y}^{l}_{m,.1}, ..., \mathbf{Y}^{l}_{m,.j-1}, \mathbf{Y}^{l-1}_{m,.j+1}, ..., \mathbf{Y}^{l-1}_{m,.J}, \{\mathbf{Y}_{m,ij}\}^{I_j}_{i=1}, \mathbf{Y}_{o}) \\
\vdots & & \\
\text{Draw}\mathbf{Y}^{l}_{m,.J} & \sim & f_{\boldsymbol{\theta}}(\mathbf{Y}_{m,.J}|\mathbf{Y}^{l}_{m,.1}, ..., \mathbf{Y}^{l}_{m,.J-1}, \{\mathbf{Y}_{m,ij}\}^{I_J}_{i=1}, \mathbf{Y}_{o})
\end{array} \qquad (4.5)
$$

In the same manner, the Gibbs sampler is coupled with the MH algorithm. Once the level-1 candidate draws are accepted or rejected, level-2 random effects candidates are generated as $\mathbf{Y}^{*}_{m,.j} = \mathbf{Y}_{m,.j} + \mathbf{e}_{.j}$. Similarly, $\mathbf{e}_{.j}$ is drawn from a scaled standard multivariate normal distribution $N_3(0, w^2 I_3)$, Conditional on $\mathbf{Y}_{m,ij}$, now level-2 draws are generated and evaluated in the same manner. The only difference is that the likelihoods are evaluated at level-2 as $\mathbf{Y}_{m,.j}$ are level-2 random effects and latent variables. For this process after simplification, the acceptance probability of moving from state $\mathbf{Y}_{m,.j}$ to $\mathbf{Y}^{*}_{m,.j}$ is calculated as follows:

$$
\begin{aligned}
& \alpha(\mathbf{Y}_{m,.j}, \mathbf{Y}^{*}_{m,.j}|\theta, \mathbf{Y}_{o,ij}, \mathbf{Y}_{m,ij}) \\
= \min & \left\{ \frac{\Pi^{I_j}_{i=1} f_{\theta}(\mathbf{Y}_{o,ij}|\mathbf{Y}^{*}_{m,ij}, \mathbf{Y}^{*}_{m,.j}) h_2(\mathbf{Y}^{*}_{m,.j}|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)}{\Pi^{I_j}_{i=1} f_{\theta}(\mathbf{Y}_{o,ij}|\mathbf{Y}_{m,ij}, \mathbf{Y}_{m,.j}) h_2(\mathbf{Y}_{m,.j}|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)}, 1 \right\}
\end{aligned} \qquad (4.6)
$$

By alternating sampling level-1 missing data conditional on level-2 missing data and sampling level-2 missing data conditional on level-1 missing data, the MH sampler makes the sequence of drawings converge in distribution to $F_{\boldsymbol{\theta}}(\mathbf{Y}_m|\mathbf{Y}_o)$ (Gelfand & Smith, 1990; Geman & Geman, 1984).

4.1.1 Tuning Constants

Recall from Equation (4.6), the distributions of level-1 and level-2 missing data are defined by $\boldsymbol{\mu}_{1j}$, $\boldsymbol{\Sigma}_{1j}$, $\boldsymbol{\mu}_2$, and $\boldsymbol{\Sigma}_2$, and they are presented in Table 4.1. At level 1, δ_{ij} and r_{ij} are treated as missing values and the mean vector is fixed at zero vector. This is because δ_{ij} is the individual deviations from group mean, and the group mean is also centered on grand mean. The expected within level residuals follow a standard normal with variance 1. As β_{0j} and β_{1j} are the combinations of fixed effects and random components u_{0j} and u_{1j}, drawing β_{0j} and β_{1j} is basically the same as drawing u_{0j} and u_{1j}. The means of these two random components are all zeros. Therefore, $\boldsymbol{\mu}_{1j}$ = (0,0)′ for both contextual effect model and a cross-level interaction model, and $\boldsymbol{\mu}_2$ = (0,0)′ and (0,0,0)′ for each model, respectively. The variance-covariance matrix at level 2 is defined as we parameterized in the models, which is the τ matrix in the traditional HLM framework.

To determine the size of tuning constant w that yields between 20 to 30% of acceptance rate at each level (Gelman, Gilks, & Roberts, 1997), an experiment was conducted in which the tuning constants were varied. Based on the results summarized in Table 4.2, the combination of 1.2 and 0.2 was chosen for a compositional effect model. For a cross-level interaction model, the combination of 1.2 and 0.12 was used since this model is a higher dimensional model with one more random effect. As Cai (2008) suggested, high dimensional model requires a much smaller tuning constant particulary for level 2. It is because the group level values are aggregated and much more stable and naturally a smaller tuning constant is needed to obtain targeted acceptance rates.

4.1.2 "Burn-in"

Another condition that should be determined in implementing an MH sampler is how the "burn-in" process should be made. Examination of auto correlations of random drawings and monitoring the traces of parameter estimates can provide needed information to make such decisions. The time series plots for every random effect drawings at level 1 and level 2 for a simulated data set with 2000 individuals nested in 100 groups are reported in Appendix A. The plots suggest that at least 20 burn-in cycles is preferable. Therefore, the number of burn-in cycles was initially set at 20. However, when the sampler is combined with the RM update for parameter estimates, further examination of the traces of parameter estimates was conducted. After comparisons of the traces from 20 burn-in cycles and 5 burn-in cycles(see section 4.3), the final burn-in cycle was decided as 5 and this number was used throughout this study.

4.2 Complete Data Models and Derivatives for Stochastic Approximation and the RM Update

With the imputations that the MH sampler generates, $\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}|\mathbf{Y}_o)$ in Equation (3.2) can be approximated as the second step of MH-RM algorithm, stochastic approximation. As it is described in Equations (3.7) and (3.8), the sample average of complete data gradients and the conditional expected distribution of missing data given observed data are calculated. Finally, the third step of the MH-RM algorithm, RM update is made by the Equation (3.9) with a set of gain constants. The iterations of these three steps converge to the MLE. As the complete data log-likelihood $l(\boldsymbol{\theta}|\mathbf{Y})$ and its derivatives $\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}|\mathbf{Y})$ are needed for Equations (3.7), (3.8) and (3.9), the first and second order derivatives of the complete data models with respect to unrestricted parameters are described in the following subsections.

4.2.1 Latent Structure Models

Denote the expected value and covariance matrix of $\boldsymbol{\eta}$ by $\boldsymbol{\mu}$ and Σ. When $\boldsymbol{\mu}$ and Σ contain parameter vectors $\boldsymbol{\theta}$ and $\boldsymbol{\tau}$ respectively, the complete data log-likelihood function can be written as,

$$l = -\frac{1}{2}[\boldsymbol{\eta} - \boldsymbol{\mu}(\boldsymbol{\theta})]'[\Sigma(\boldsymbol{\tau})]^{-1}[\boldsymbol{\eta} - \boldsymbol{\mu}(\boldsymbol{\theta})] - \frac{1}{2}log|\Sigma(\boldsymbol{\tau})| - \frac{1}{2}Nlog2\pi. \tag{4.7}$$

Then the first derivative of l with respect to the parameter vector $\boldsymbol{\theta}$ is

$$\frac{\partial l}{\partial \boldsymbol{\theta}} = \frac{\partial \boldsymbol{\mu}'}{\partial \boldsymbol{\theta}} \Sigma(\boldsymbol{\tau})^{-1}(\boldsymbol{\eta} - \boldsymbol{\mu}(\boldsymbol{\theta})). \tag{4.8}$$

The first derivative of l with respect to a parameter τ_k is

$$\frac{\partial l}{\partial \tau_k} = -\frac{1}{2}\left[tr(\Sigma^{-1}\frac{\partial \Sigma}{\partial \tau_k}) - (\boldsymbol{\eta} - \boldsymbol{\mu})'\Sigma^{-1}\frac{\partial \Sigma}{\partial \tau_k}\Sigma^{-1}(\boldsymbol{\eta} - \boldsymbol{\mu})\right]. \tag{4.9}$$

The second derivative of l with respect to the parameter vector $\boldsymbol{\theta}$ is

$$\frac{\partial^2 l}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = -\frac{\partial \boldsymbol{\mu}'}{\partial \boldsymbol{\theta}}\Sigma^{-1}\frac{\partial \boldsymbol{\mu}'}{\partial \boldsymbol{\theta}'} + \left\{(\boldsymbol{\eta} - \boldsymbol{\mu})'\Sigma^{-1}\frac{\partial^2 \boldsymbol{\mu}}{\partial \theta_i \partial \boldsymbol{\theta}'}\right\}. \tag{4.10}$$

The second derivative of l with respect to parameters τ_k and τ_s is

$$\begin{aligned}\frac{\partial^2 l}{\partial \tau_s \partial \tau_k} = & -\frac{1}{2}\Big\{tr\Big(\Sigma^{-1}\frac{\partial \Sigma}{\partial \tau_s}\Sigma^{-1}\frac{\partial \Sigma}{\partial \tau_k}\Sigma^{-1}\frac{\partial^2 \Sigma}{\partial \tau_s \partial \tau_k}\Big) \\ & + (\boldsymbol{\eta} - \boldsymbol{\mu})'\Big[(-1)\Sigma^{-1}\frac{\partial \Sigma}{\partial \tau_s}\Sigma^{-1}\frac{\partial \Sigma}{\partial \tau_k}\Sigma^{-1} + \Sigma^{-1}\frac{\partial^2 \Sigma}{\partial \tau_s \partial \tau_k}\Sigma^{-1} \\ & - \Sigma^{-1}\frac{\partial \Sigma}{\partial \tau_k}\Sigma^{-1}\frac{\partial \Sigma}{\partial \tau_s}\Sigma^{-1}\Big](\boldsymbol{\eta} - \boldsymbol{\mu})\Big\}. \end{aligned} \tag{4.11}$$

4.2.2 Graded Responses

For the manifest variables that have more than two categories, Equation (2.9) can be redefined as follows, suppressing subscripts:

$$\begin{aligned} T_0 &= 1, \\ T_1 &= \frac{1}{1 + \exp[-(b_{1,l} + a\xi)]}, \\ T_2 &= \frac{1}{1 + \exp[-(b_{2,l} + a\xi)]}, \\ &\vdots \\ T_{K-1} &= \frac{1}{1 + \exp[-(b_{K_l-1,l} + a\xi)]}, \\ T_K &= 0 \end{aligned}$$

The cumulative response probability for a category k is defined as $P_k = T_k - T_{k+1}$. Taking the log of the likelihood function of the complete data model yields the following equation,

$$l = \sum_{k=0}^{K-1} \chi_k(x) log P_k = \sum_{k=0}^{K-1} \chi_k(x) log(T_k - T_{k+1}), \tag{4.12}$$

where x is the response to a graded item with K categories. The first derivatives of the complete data model log-likelihood are

$$\begin{aligned} \frac{\partial l}{\partial b_k} &= \frac{\partial}{\partial b_k}(\chi_{k-1}(x) log(T_{k-1} - T_k) + \chi_k(x) log(T_k - T_{k+1})) \\ &= -(\frac{\chi_{k-1}(x)}{T_{k-1} - T_k} - \frac{\chi_k(x)}{T_k - T_{k+1}})\frac{\partial T_k}{\partial b_k} \\ \frac{\partial l}{\partial a} &= \sum_{k=0}^{K-1} \frac{\chi_k(x)}{T_k - T_{k+1}}(\frac{T_k}{\partial a} - \frac{T_{k+1}}{\partial a}), \end{aligned}$$

where

$$\frac{\partial T_k}{\partial b_k} = T_k(1 - T_k), \frac{\partial T_k}{\partial a} = T_k(1 - T_k)\xi.$$

The second derivatives are given by

$$
\begin{aligned}
\frac{\partial^2 l}{\partial b_k^2} &= -(\frac{\chi_{k-1}(x)}{(T_{k-1}-T_k)^2}+\frac{\chi_k(x)}{(T_k-T_{k+1})^2})(\frac{\partial T_k}{\partial b_k})^2 \\
&\quad -(\frac{\chi_{k-1}(x)}{T_{k-1}-T_k}-\frac{\chi_k(x)}{T_k-T_{k+1}})(\frac{\partial}{\partial b_k}\frac{\partial T_k}{\partial b_k}) \\
\frac{\partial^2 l}{\partial b_{k-1}\partial b_k} &= \frac{\chi_{k-1}(x)}{(T_{k-1}-T_k)^2}(\frac{\partial T_{k-1}}{\partial b_{k-1}})(\frac{\partial T_k}{\partial b_k}) \\
\frac{\partial^2 l}{\partial b_{k+1}\partial b_k} &= \frac{\chi_k(x)}{(T_{k+1}-T_k)^2}(\frac{\partial T_{k+1}}{\partial b_{k+1}})(\frac{\partial T_k}{\partial b_k}) \\
\frac{\partial^2 l}{\partial a\partial b_k} &= -\frac{\chi_k(x)}{(T_{k+1}-T_k)^2}(\frac{\partial T_k}{\partial b_k})(\frac{\partial T_k}{\partial a}-\frac{\partial T_{k+1}}{\partial a}) \\
&\quad +\frac{\chi_{k-1}(x)}{(T_{k-1}-T_k)^2}(\frac{\partial T_k}{\partial b_k})(\frac{\partial T_{k-1}}{\partial a}-\frac{\partial T_k}{\partial a}) \\
&\quad -(\frac{\chi_{k-1}(x)}{T_{k-1}-T_k}-\frac{\chi_k(x)}{T_k-T_{k+1}})(\frac{\partial}{\partial a}\frac{\partial T_k}{\partial b_k}) \\
\frac{\partial^2 l}{\partial a\partial a'} &= \sum_{k=0}^{K-1}\{-\frac{\chi_k(x)}{(T_k-T_{k+1})^2}(\frac{\partial T_k}{\partial a}-\frac{\partial T_{k+1}}{\partial a})(\frac{\partial T_k}{\partial a'}-\frac{\partial T_{k+1}}{\partial a'}) \\
&\quad +\frac{\chi_k(x)}{T_k-T_{k+1}}(\frac{\partial}{\partial a}\frac{\partial T_k}{\partial a'}-\frac{\partial}{\partial a}\frac{\partial T_{k+1}}{\partial a'})\},
\end{aligned}
$$

where

$$
\begin{aligned}
\frac{\partial}{\partial b_k}\frac{\partial T_k}{\partial b_k} &= T_k(1-T_k)(1-2T_k) \\
\frac{\partial}{\partial a}\frac{\partial T_k}{\partial b_k} &= T_k(1-T_k)(1-2T_k)\xi \\
\frac{\partial}{\partial a}\frac{\partial T_k}{\partial a} &= T_k(1-T_k)(1-2T_k)\xi\xi'.
\end{aligned}
$$

4.3 Acceleration and Convergence

Asymptotically in time, the MH-RM algorithm converges to the MLE. However, Cai (2008) pointed out that the algorithm can be stuck in locations that are not close to the MLE during the initial stage of iterations due to the sequence of gain constants being deterministic and eventually going to zero. In this case, premature convergence can occur. As a solution, using adaptive gain constants is suggested in a three-stage gain procedure (Cai, 2008). The first stage is initial M1 iterations with constant gain constant 1. At the end of iteration M1, run another M2 iterations. The parameter estimates that are updated during the M2 iterations are averaged and used as starting values again in the final stage of iterations with decreasing gain constants. The three stage procedure ensures that the algorithm can effectively and stably converge to the MLE.

As a convergence check method, Cai (2008) proposed to monitor a "window" of the largest difference between two successive parameter estimates, in which the iterations stop when all of monitored differences are less than a small number. Cai (2008) suggested 3 as a reasonable width of the window to be monitored in practice.

To find proper conditions of number of iterations, magnitude of gain constants, and convergence criteria, the traces of parameter estimates were examined. Figures 4.1, 4.2, 4.3, 4.4, and 4.5 are the time series plots of all measurement and statical parameter estimates when every 20th random drawing was used to approximate the score vectors and the hessian matrix and update parameter estimates when a constant gain was 0.1 and the decreasing gain constant γ_k at kth iteration is defined as,

$$\gamma_k = \frac{0.1}{k^{\epsilon}}. \tag{4.13}$$

The value of ϵ was 0.75 for this study after examining the traces of estimates .

To see the behavior of the parameter estimates as the iterations proceed, a large number of iterations (1000) was used and another 1000 iterations were attempted for the decreasing gain constant stage. The plots suggest that at least 100 iterations are needed for the initial M1 stage to let the parameter estimates move to close to the MLEs. Then about 300 to 500 iterations are enough to see that the estimates are oscillating around the MLEs. When 1.0×10^{-5} was used as the convergence criteria, 1,000 iterations for decreasing gain constant stage (M3) was not enough and the iteration does not stop. When 1.0×10^{-4} was used as the criteria, the iteration stops but the point estimates are still slightly different from MLEs that are obtained from the EM algorithm. Accordingly,

0.5×10^{-5} which is between 1.0×10^{-5} and 1.0×10^{-4} was used for the convergence check for this study. This means that when the largest difference among three successive parameter estimates gets smaller than 0.5×10^{-5}, the iteration stops. The reported time-series plots show that after about 500 iterations at the decreasing gain constants stage, the sequence of parameter estimates satisfy the convergence criterion.

In addition, under the same conditions, 5 burn-in cycles were tried and the time series plots of parameter estimates are reported in Figures 4.6, 4.7, 4.8, 4.9, and 4.10. As the time series plots of the parameter estimates appear to be similar to those when 20 burn-in cycles were used, 5 burn-in cycles were chosen for higher efficiency throughout the study.

Table 4.1: Latent variable distributions

		Contextual Effect Model		
Level 1	$\begin{bmatrix} \delta_{ij} \\ r_{ij} \end{bmatrix}$	$\sim$	$\boldsymbol{\mu}_{1j} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$	$\boldsymbol{\Sigma}_{1j} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$
Level 2	$\begin{bmatrix} \xi_{.j} \\ u_{0.j} \end{bmatrix}$	$\sim$	$\boldsymbol{\mu}_{2} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$	$\boldsymbol{\Sigma}_{2} = \begin{bmatrix} \tau_{00} & 0 \\ 0 & Var(\xi_{.j}) \end{bmatrix}$
		Cross-level Interaction Model		
Level 1	$\begin{bmatrix} \delta_{ij} \\ r_{ij} \end{bmatrix}$	$\sim$	$\boldsymbol{\mu}_{1j} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$	$\boldsymbol{\Sigma}_{1j} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$
Level 2	$\begin{bmatrix} \xi_{.j} \\ u_{0.j} \\ u_{1.j} \end{bmatrix}$	$\sim$	$\boldsymbol{\mu}_{2} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$	$\boldsymbol{\Sigma}_{2} = \begin{bmatrix} \tau_{00} & \tau_{01} & 0 \\ \tau_{10} & \tau_{11} & 0 \\ 0 & 0 & Var(\xi_{.j}) \end{bmatrix}$

Table 4.2: Tuning constants and acceptance rates of drawings

	Set 1		Set 2		Set 3		Set 4	
	L1	L2	L1	L2	L1	L2	L1	L2
Tuning constant	1	1	1.2	0.8	1.2	0.5	1.2	0.17
Acceptance rate (%)	34-38	3-4	30-36	4-6	32-35	6-10	32-35	20-32

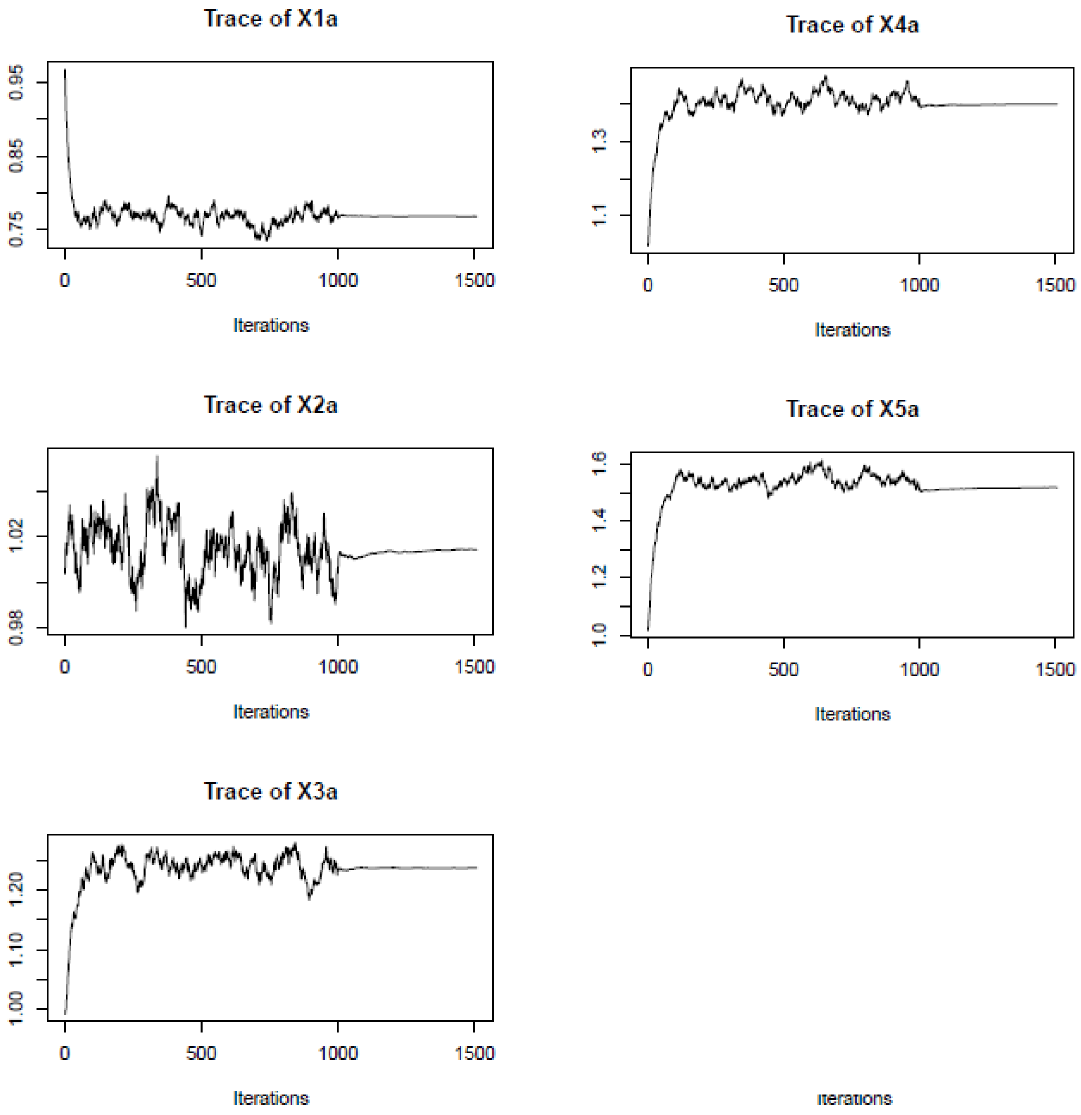


Figure 4.1: The time-series plots of slope estimates for X side manifest variables (total sample size=$2,000$, number of groups=100, compositional effects model simulated data, 20 burn-in cycles)

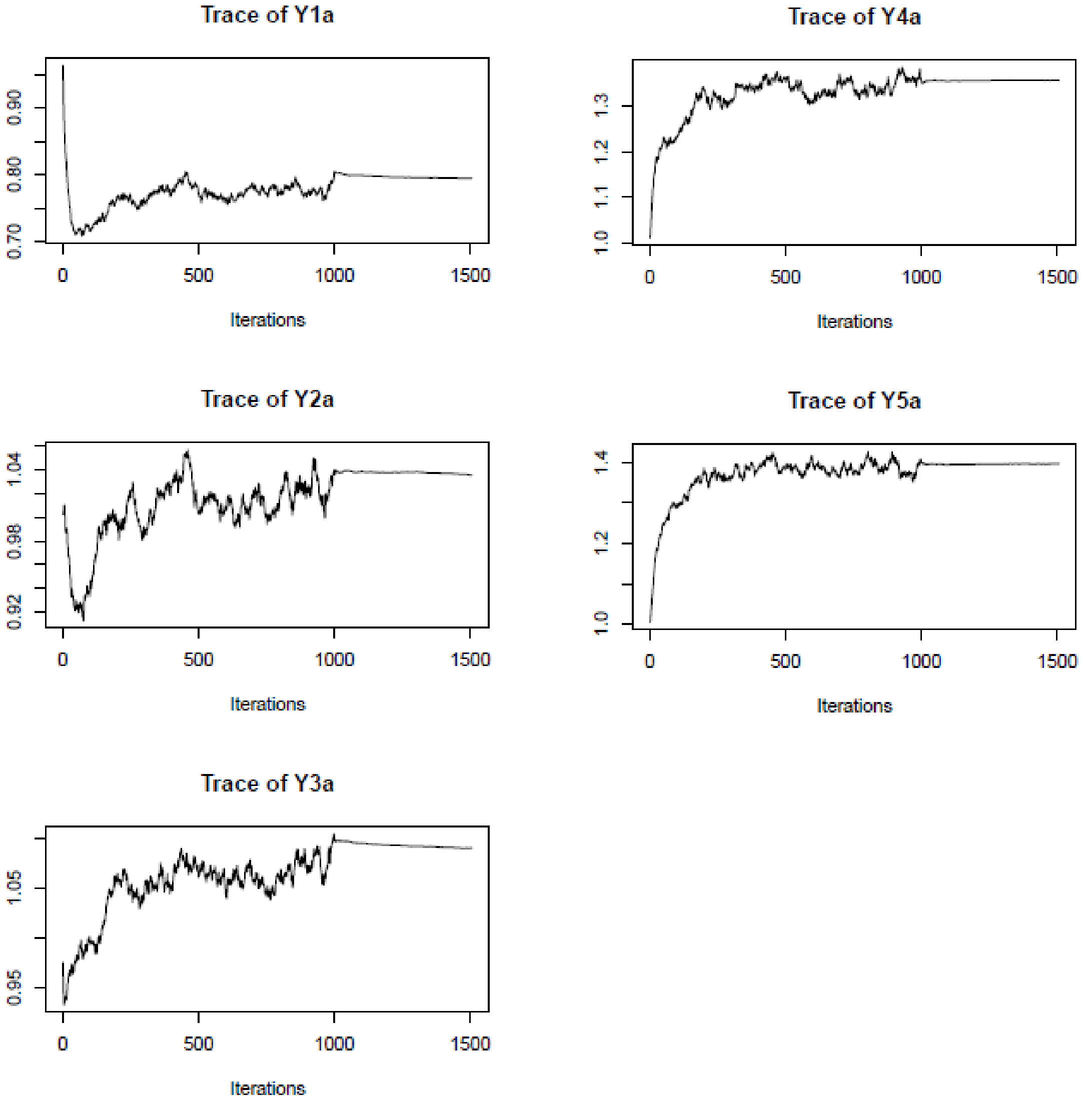


Figure 4.2: The time-series plots of slope estimates for Y side manifest variables (total sample size=2,000, number of groups=100, compositional effects model simulated data, 20 burn-in cycles)

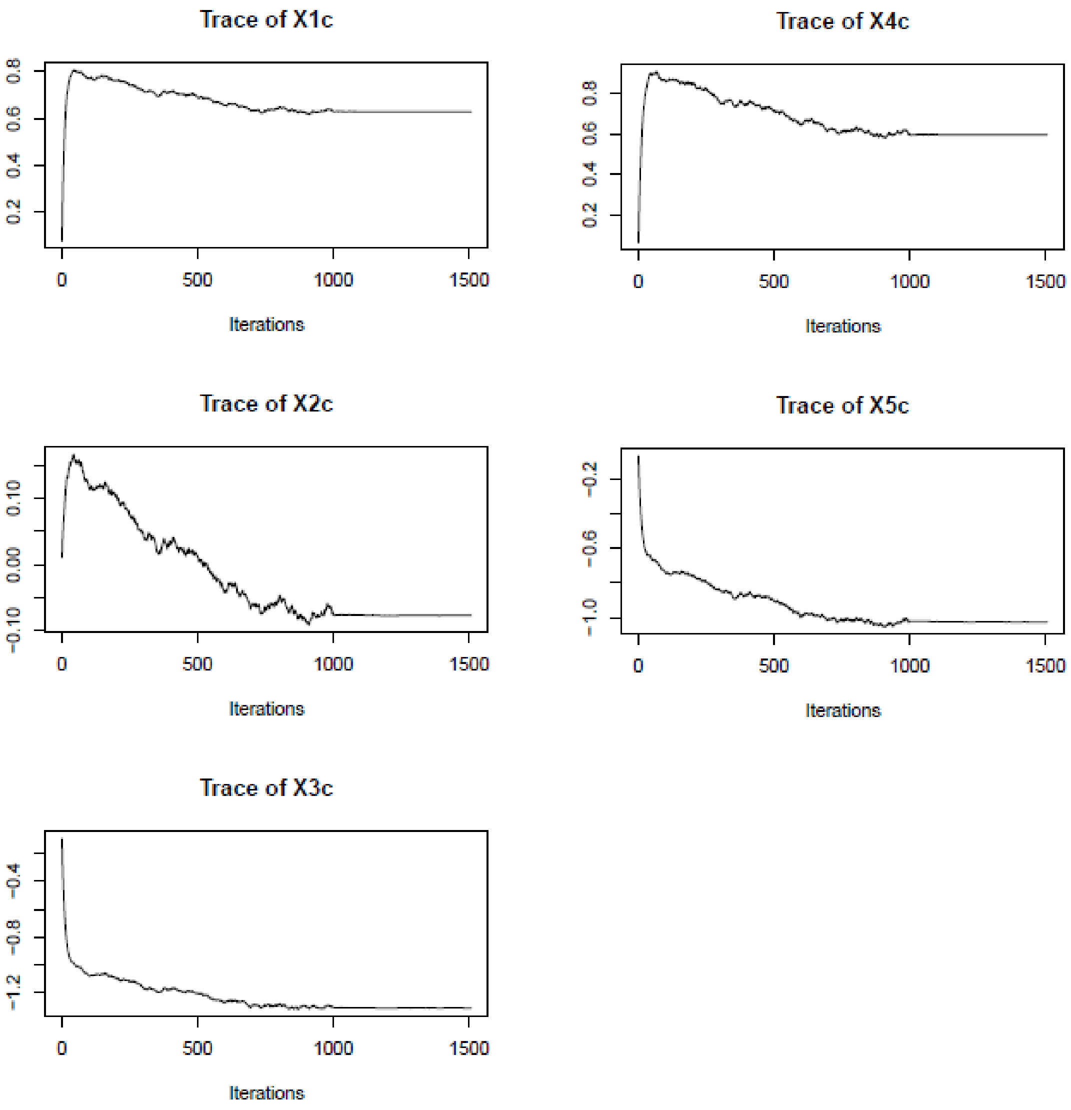


Figure 4.3: The time-series plots of intercept estimates for X side manifest variables (total sample size=2,000, number of groups=100, compositional effects model simulated data, 20 burn-in cycles)

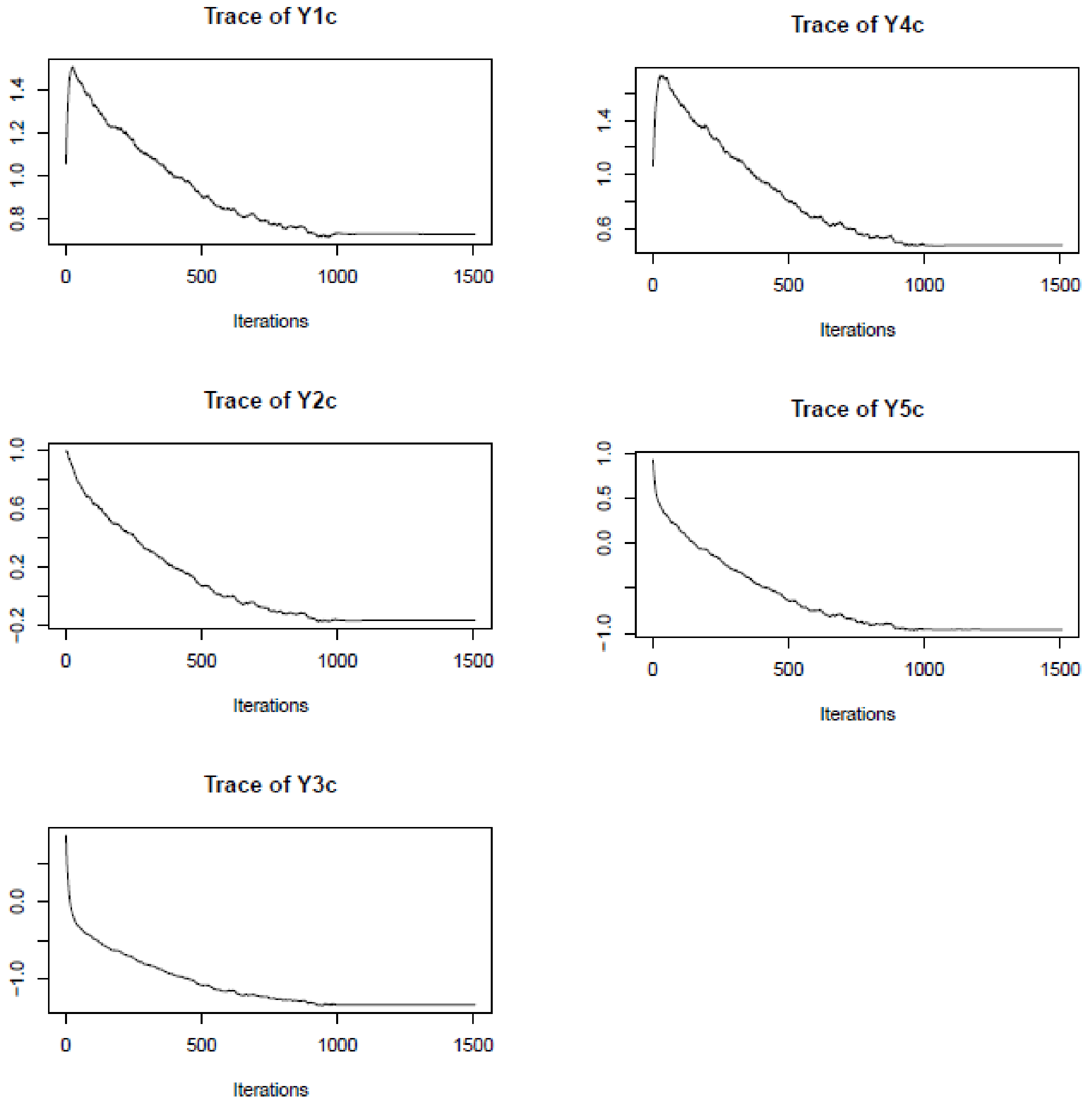


Figure 4.4: The time-series plots of intercept estimates for Y side manifest variables (total sample size=$2,000$, number of groups=100, compositional effects model simulated data, 20 burn-in cycles)

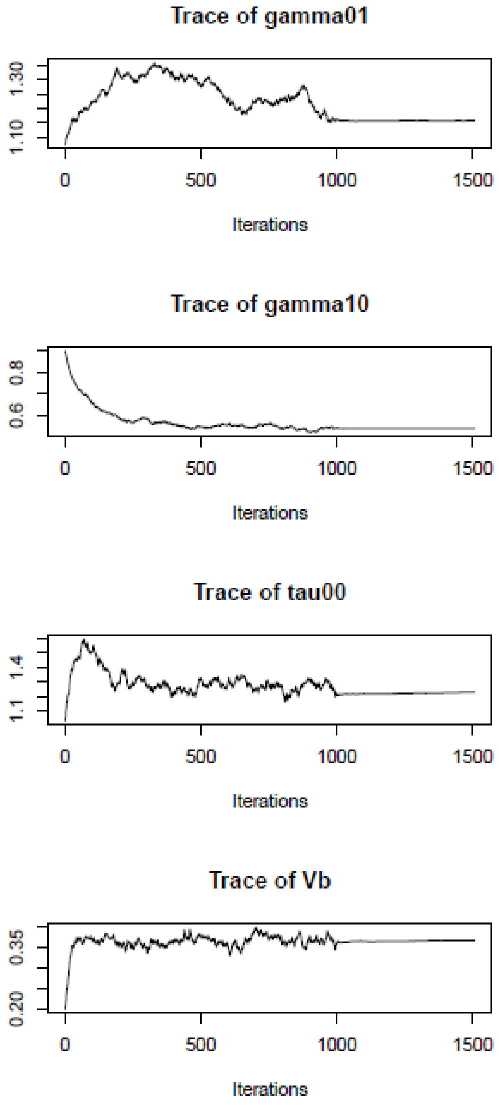


Figure 4.5: The time-series plots of structural parameter estimates (total sample size=2, 000, number of groups=100, compositional effects model simulated data, 20 burn-in cycles)

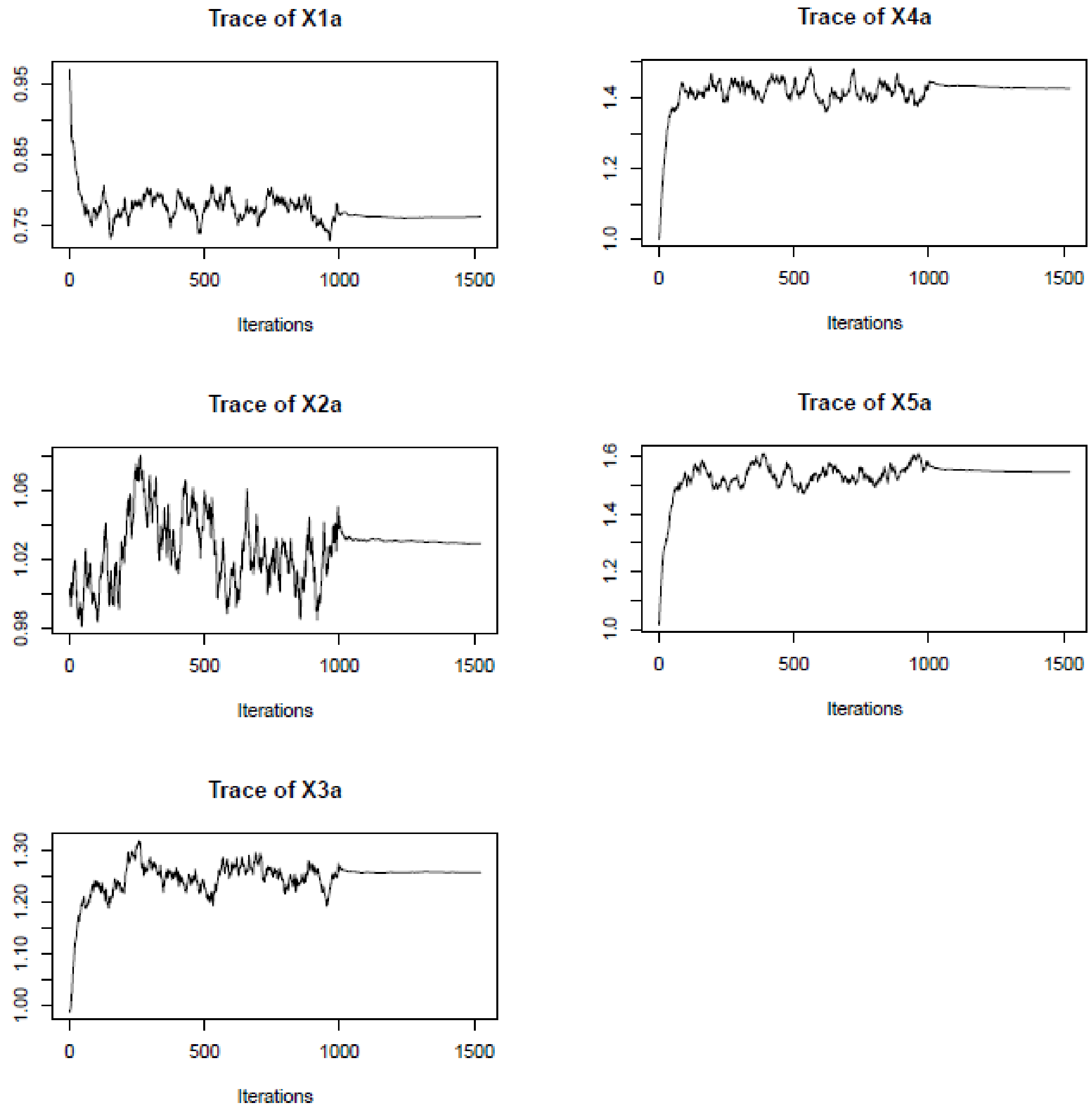


Figure 4.6: The time-series plots of slope estimates for X side manifest variables (total sample size=2, 000, number of groups=100, compositional effects model simulated data, 5 burn-in cycles)

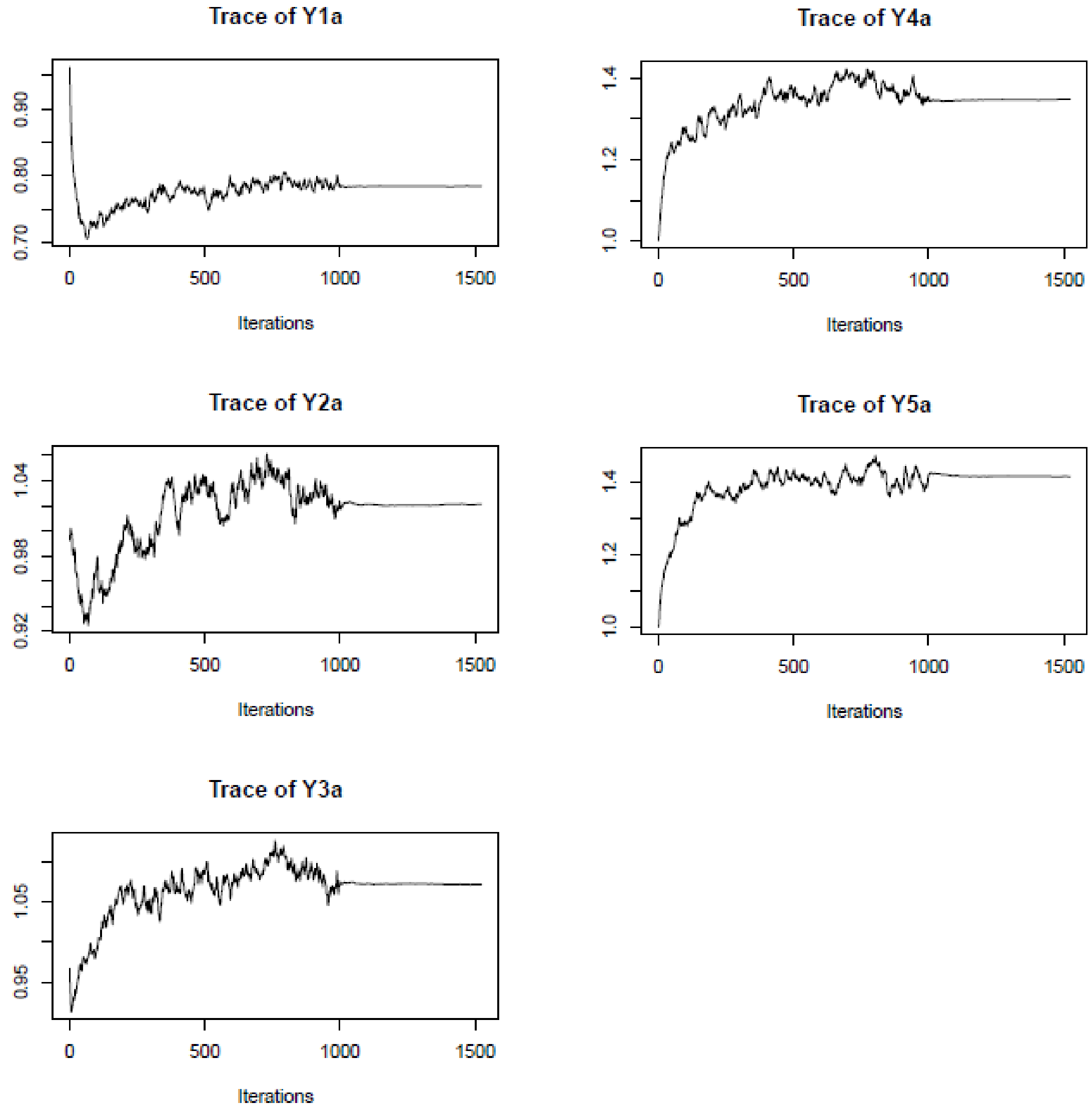


Figure 4.7: The time-series plots of slope estimates for Y side manifest variables (total sample size=$2,000$, number of groups=100, compositional effects model simulated data, 5 burn-in cycles)

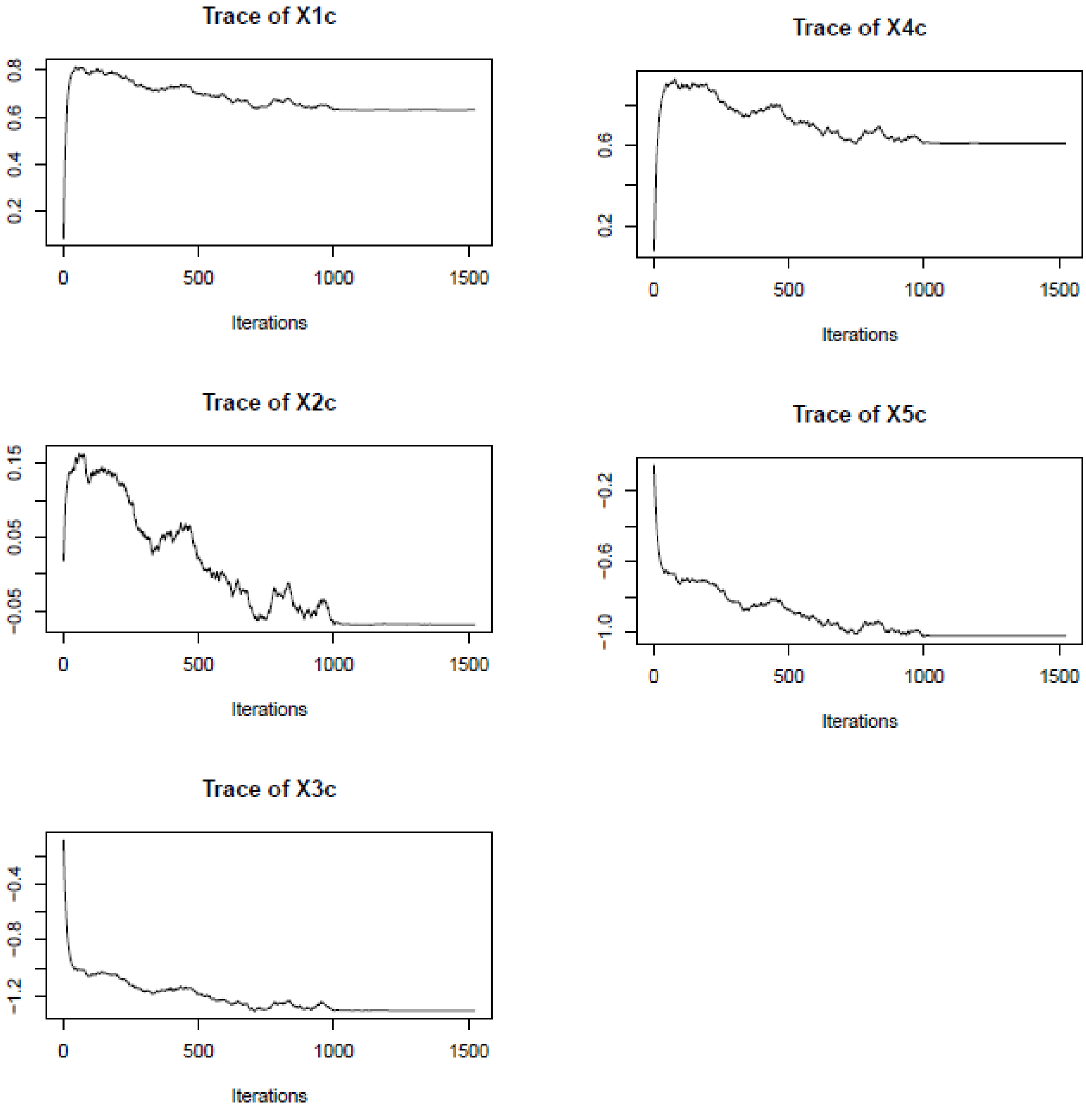


Figure 4.8: The time-series plots of intercept estimates for X side manifest variables (total sample size=2, 000, number of groups=100, compositional effects model simulated data, 5 burn-in cycles)

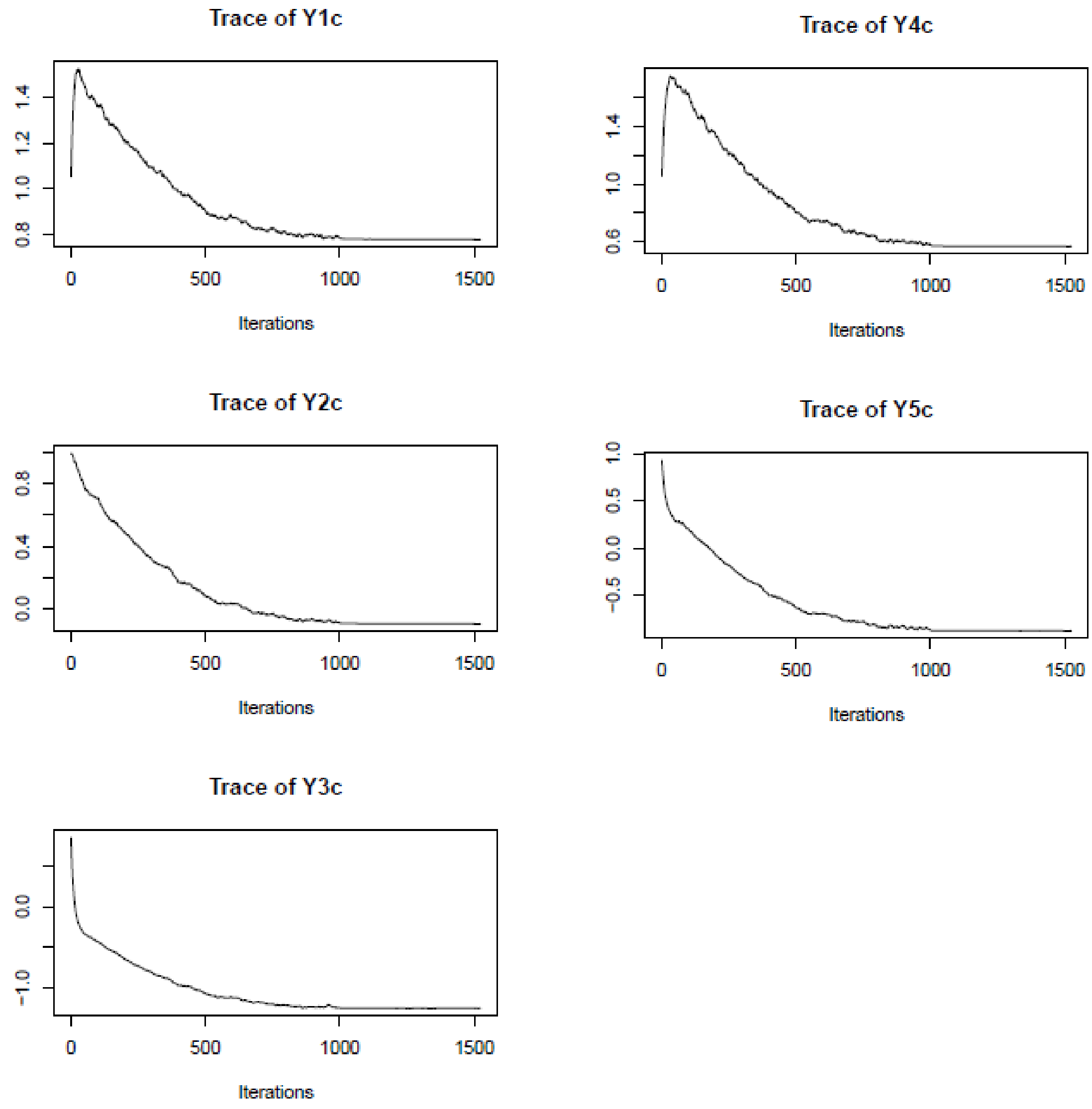


Figure 4.9: The time-series plots of threshold estimates for Y side manifest variables (total sample size=2,000, number of groups=100, compositional effects model simulated data, 5 burn-in cycles)

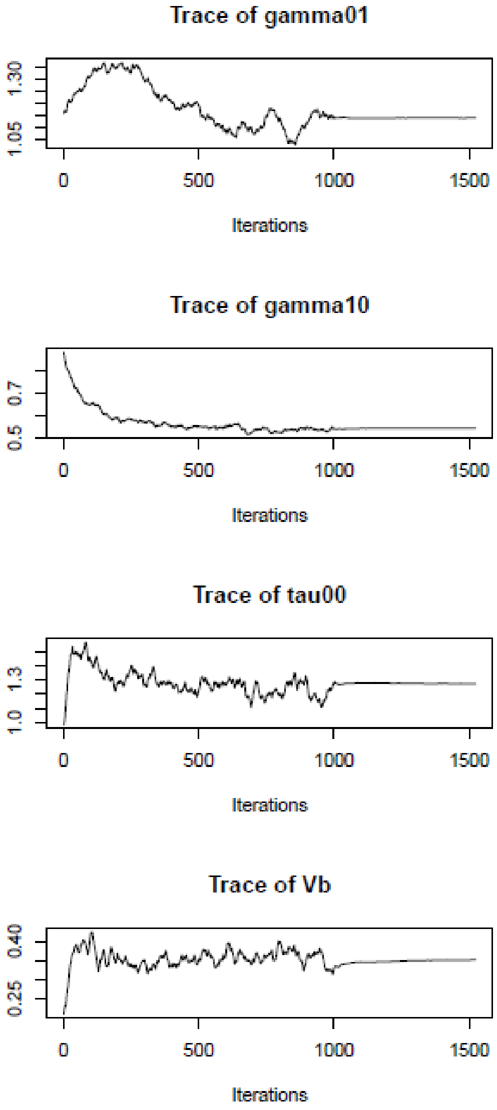


Figure 4.10: The time-series plots of structural parameter estimates (total sample size=2,000, number of groups=100, compositional effects model simulated data, 5 burn-in cycles)

CHAPTER 5

Simulation Studies

Two simulation studies were conducted with two distinct objectives. The first study examined the parameter recovery and standard errors when MH-RM algorithm is implemented in comparison to those from an existing EM algorithm. The aim of the second study was to compare the performance of estimating a compositional effect and a cross-level interaction between a traditional HLM model that ignores measurement and sampling error and the multilevel latent variable model that takes the two error sources into account. This chapter summarizes the methods and results of the two simulation studies.

5.1 Simulation Study 1: Comparison of Estimation Algorithms

5.1.1 Methods

To examine parameter recovery and standard errors between the MH-RM algorithm and EM algorithm, simulated data were generated under a favorable sampling condition with a simple measurement structure. Here, a favorable sampling condition means large sizes at both levels and a sufficiently large ICC for the predictor latent variable based on previous research.

The data-generating and fitted models followed Equation (2.1) for a compositional effect model and Equation (2.4) for a cross-level interaction model. The simulated data are balanced in that the number of level-2 units (*ng*) is 100 and the number of level-1 units per group (*np*) is 20. The generating ICC value for the latent predictor was 0.3. For the measurement model, five dichotomously scored manifest variables were generated for each latent trait (i.e., η, and ξ) using a 2-PL model. For η_{ij}, the manifest variables are Y_1, Y_2, Y_3, Y_4, and Y_5. For ξ_{ij}, which is the sum of deviations and the level-2 latent mean ($\delta_{ij} + \xi_{.j}$), the manifest variables are X_1, X_2, X_3, X_4, and X_5. The item parameters were the same across levels, representing cross-level measurement invariance.

100 data sets were generated with the same parameters but with 100 different random seeds for each model. The first 10 data sets were analyzed using two methods: an MH-RM algorithm implemented in R (R Core Team, 2012) and an adaptive quadrature EM approach implemented in Mplus (Muthén & Muthén, 2010). Then the other 90 data sets are all analyzed using the MH-RM algorithm. Standardized summed scores for the predictor and outcome manifest variables were used as starting values for ξ_{ij} and η_{ij}. Fixed parameter starting values were obtained from traditional HLM model estimates using these standardized summed scores. Starting values for level-1 and level-2 random variable samples are randomly drawn from a standard normal distribution.

For compositional effect estimation, the MH-RM algorithm's convergence criterion was 5.0×10^{-5}, and the maximum numbers of iterations for each stage were $M1 = 100$, $M2 = 500$, and $M3 = 600$. For the cross-level interaction model, the MH-RM algorithm convergence criterion was 5.0×10^{-5} and the maximum numbers of iterations for each stage were $M1 = 100$, $M2 = 800$, and $M3 = 800$. To calculated post-convergence approximated standard errors, 100 to 500 samples were used for the compositional effect model, and 100 to 800 samples were used for the cross-level interaction model. The convergence rates at the given number of iterations were 100% and 52% for the compositional effect model and the cross-level interaction model, respectively.

5.1.2 Results

5.1.2.1 Compositional Effect Model

The following results indicate that a compositional effect model can be efficiently estimated through the MH-RM algorithm, requiring less time than an EM algorithm with 14 adaptive quadrature points. All point estimates in the measurement and structural model were reasonably close to generating values. While post-convergence approximation of observed information method was more efficient than the recursive approximation, the standard errors of item intercepts appeared to be smaller when the former method is applied.

The generating values and the corresponding estimates for the compositional effect model from different algorithms are summarized in Table 5.1. The first column contains the true parameters for the measurement and structural parameters. The second set of columns and the third set of columns include the estimates and SEs from EM with different numbers of adaptive quadrature points (5 and 14). The default number of quadrature points is 15 in Mplus, but the computer memory cannot handle 15 quadrature points for this four-dimensional model. The maximum possible number of quadrature points was 14 for a compositional effect model. A smaller number of quadrature points (5) was tested to compare the results in terms of point estimates and standard errors of estimation. The fourth set of columns includes the corresponding values using the MH-RM algorithm.

The means of point estimates from different algorithms are generally very close to one another. For structural parameter estimates in the first panel, the number of quadrature points does not appear to make a large difference, though 14-quadrature-point estimates are slightly closer to the MH-RM estimates and the generating values in terms of τ_{00} and $var(\xi_{.j})$. Standard errors are also very similar.

For measurement parameter estimates, both the means of point estimates and the standard errors were the same up to the second decimal point across different numbers of quadrature points. The biggest difference in means of point estimates between EM algorithm and the MH-RM algorithm was 0.02, indicating that the two approaches yield roughly identical estimates. However, mean standard error estimates are slightly different between MH-RM and EM results in that the standard error estimates from MH-RM algorithm for intercepts are smaller than those form EM algorithm. The biggest difference in standard error estimates for measurement parameters between two algorithms was 0.13. This may be due to the difference in SE calculation across programs.

The log of standard error estimates from EM algorithm and log of post-convergence approximated standard errors from MH-RM algorithm are plotted against log standard deviations of point estimates in Figure 5.1. The estimates are clustered on the diagonal line, indicating that estimated standard errors are generally close to the Monte Carlo standard deviations of the point estimates, except for the intercept parameter standard errors, which appear to be underestimated when the post-convergence approximation is used for the MH-RM algorithm.

When one processor was used for estimation, 5 quadrature point EM required a very short time, while 14 quadrature point EM required over an hour. The MH-RM algorithm required about 40 minutes. Note that MH-RM is implemented in R (an interpreted language), while Mplus is written in FORTRAN (a compiled language). As an interpreted language is expected to be slow compared to a compiled language, a direct comparison is inappropriate.

To examine the performance of the MH-RM algorithm further, 100 generated data sets were analyzed, and the results are summarized in Table 5.2. The means of point estimates are reasonably close to generating values in general, with slight underestimation of variance estimates in the structural parameters. For structural parameters, the Monte Carlo standard deviations of parameter estimates (column 5) are also similar to both standard error estimates (column 4 and 6); the largest difference is 0.02. With respect to point estimates, means of item parameter estimates are very close to generating values.

However, recursively approximated standard errors are closer to the Monte Carlo standard deviations of item parameter estimates than the post-convergence approximated standard errors. More specifically, the most prominent differences are found in the standard errors of intercept parameters in that post-convergence approximated standard errors for item intercept parameters are underestimated. Therefore, recursively approximated standard errors perform better than post-convergence approximated standard errors. However, a drawback of using recursively approximated standard errors is the requirement of a large number of iterations (at least 1000). In addition, given the prespecified maximum 1,500 M3 iterations, only half of the replications reached a properly converged (i.e., positive definite) observed data information matrix. For this reason, calculation of post-convergence approximated standard errors is adopted for the remaining simulations in this study since this approach gives proper standard error estimates in general, except for those that are associated with item intercept parameters.

Finally, 95% confidence intervals of each parameter estimate using the post-convergence approximated standard errors were calculated. The percentages of intervals that cover the generating values are reported in the last column of Table 5.2. Based on the 100 replications performed, coverage of fixed structural parameters appears proper, in general. For measurement parameters, the coverage rates tend to be lower as the magnitude of parameter gets larger. Coverage rates are the lowest for the large threshold parameters due to the underestimated standard errors.

5.1.2.2 Cross-level Interaction Model

The following results indicate that a cross-level model can be efficiently estimated through the MH-RM algorithm, requiring less time than an EM algorithm with 8 adaptive quadrature points. All point estimates in the measurement and structural model were reasonably closed to generating values. The underestimated standard errors, particulary for item intercepts were consistently observed.

The generating values and the corresponding estimates from analyzing the first simulated data set using different algorithms are summarized in Table 5.3. Unlike the composition effect model results, the number of quadrature points for the EM algorithm makes some noticeable differences in the mean point estimates as well as the standard errors. The maximum possible number of quadrature points was 8 for this cross-level interaction model that requires 5-dimensional integration. The point estimates and standard errors using 8 and 5 quadrature points are reported in the first and second sets of columns in Table 5.3. The differences are particularly prominent in the structural parameters and the slopes of predictor-side indicators, as within-level variance estimates of the predictor were different across the number of quadrature points being used. However, the results from MH-RM algorithm are closer to the 8-quadrature-points results, indicating that reducing the number of quadrature points for a higher dimensional model is not desirable.

Efficiency of the MH-RM algorithm compared to the EM algorithm was more prominent for this cross-level interaction model, even as it is still in R. Using Mplus, even with 8 processors, the estimation took more than 1 hour and 30 minutes, while it took similar or even shorter time for the MH-RM algorithm implemented in R. When 1 processor was used, it took about 4 to 5 hours to yield a result using Mplus. This difference is remarkable considering that R does not have support for multi-processors.

For further analysis, more simulated data sets were analyzed by applying the MH-RM algorithm, and the generating values and corresponding estimates are summarized in Table 5.4. Results are generally similar to those obtained from the compositional effect model. The largest relative bias of the parameter estimates for both measurement and structural parts is less than 10%. Means of standard error estimates and Monte Carlo standard deviations of point estimates are reasonably compatible; however, underestimation of standard errors for threshold estimates was consistent, indicating that the post-convergence approximation approach can be chosen for efficiency reason but with a cost in accuracy.

However, only 26 of 50 replications converged within the specified number of iterations. For this condition, the cause of low convergence rate was mostly due to the approximation of observed data information matrix rather than point estimates themselves. Either allowing larger numbers of iterations or achieving more efficient approximation of the observed data information matrix would help the convergence rate increase. As a trial, 1000 iterations was tried, and this could increase the convergence rate up to 78% for this condition.

5.2 Simulation Study 2: Comparison of Models

The second simulation study was conducted to examine how measurement error and sampling error may influence compositional effect and cross-level interaction estimates across different conditions with both a traditional HLM model and a latent variable model. The methods and results are reported in following sections.

5.2.1 Methods

Simulation Conditions

A total of 42 conditions (compositional effect sizes (2) × sampling conditions (3) × ICC sizes (2) × measurement conditions (3) + no compositional effect model (6)) were examined for the compositional effect model.

First, two different sizes of compositional effect were considered in this study. The generating value of γ_{10} was either 0.5 or 0.8, giving a compositional effect of 0.5 or 0.2, respectively.

Second, the combination of large (ng=100, np=20) and small (ng=25, np=5) numbers of groups and individuals makes a total of 4 different sampling conditions. However, the combination of small number of groups and small group size leads to too small a total sample size (N=125), which is inappropriate for this kind of high-dimensional model. Therefore, only three different sampling conditions were used for this simulation study.

For latent predictor ICC levels, 0.1 and 0.3 were used to generate small- and a large-ICC conditions.

Finally, three different measurement structures were considered to address variations in measurement situations, as described in Table 5.5. The true item parameters used are reported in Table 5.6.

Additionally, data were generated from a model with no compositional effect ($\gamma_{01} = \gamma_{10}$) with the first measurement condition and analyzed to examine empirical Type I error rates for the traditional model and the latent variable model.

100 replications were attempted for each condition and the contextual effect model was applied. Similarly, two different magnitudes of the cross-level interaction effect ($\gamma_{11} = 1$ or 0.5) were considered for the cross-level interaction model, in which the

sampling, ICC, and measurement conditions are kept the same. Therefore, 54 conditions were examined for the cross-level interaction model. For the cross-interaction model, 50 replications were attempted.

Analysis

As simulated data sets have the true generating values of η_{ij} and ξ_{ij}, these values (true scores) can be analyzed using a traditional model. Then the resulting parameter estimates can be considered gold standard estimates that are influenced only by sampling fluctuations but not by measurement conditions. Therefore, each data set has three sets of parameter estimates: 1) estimates from analyzing the generating values of η_{ij} and ξ_{ij} with a traditional multilevel model, which is treated as the gold standard (denoted as G), 2) estimates obtained by applying latent variable model (denoted as L), and 3) the estimates from analyzing the observed summed scores with the manifest variable approach (denoted as M). All of the traditional HLM analyses were conducted using an R package *nlme* (Pinheiro, Bates, DebRoy, Sarkar, & R Core Team, 2012).

Statistics

To compare these three sets of estimates, four statistics are calculated: 1) the percentage bias of the estimate relative to the magnitude of generating value, 2) the root mean squared difference between the estimate and the true parameter (RMSE), 3) the observed coverage of the 95% confident interval (CI) for true value, and 4) the observed power to detect the effect of interest as significant. The percentage bias of the parameter captures how well compositional effects and cross-level interaction parameters are recovered, and RMSE captures the variability of the estimates across replications. The observed coverage of the CI can tell us how well the standard errors associated with the parameters of interest are estimated, and the observed percentage of the significant compositional effects and cross-level interaction effects can tell us how researchers' substantive decisions can be different when either a traditional model or a latent variable model is applied.

It should be noted that the regression coefficient estimates from the observed sum score analysis using a traditional multilevel model are not on the same scales as those obtained using the latent variable approach. To make the coefficient estimates more comparable, the estimates from traditional model approach were standardized by multiplying the parameter estimates by the ratio of standard deviation of the predictor to the standard deviation of the outcome.

Convergence rates and mean estimation time across generated data conditions are reported in Table 5.7. Only converged cases were taken into calculate statistics, and percentages were calculated out of converged cases. In addition, the observed coverage and power were not calculated for the cross-level interaction model with measurement models 2 and 3 because the numbers of converged cases were too small to make inferences about coverage or power.

5.2.2 Results

5.2.2.1 Compositional Effect Model

The following results indicate that the nonlinear multilevel latent variable modeling is preferred to the traditional HLM based on the examinations of bias, RMSE, coverage rates, and Type I error rates because the point estimates and standard errors are less biased and more precise across sampling conditions when the latent variable model was applied to analyze the simulated data. Particularly, substantial Type I error rates were concerned when the traditional HLM was applied.

Relative percentage bias

Since a compositional effect estimate is defined as the difference between $\hat{\gamma}_{01}$ and $\hat{\gamma}_{10}$, those parameter estimates are examined together, along with the compositional effect estimates. Relative percentage bias of these three estimates across data conditions and models are summarized in Tables 5.8, 5.9, and 5.10. The first panel is for relative percentage bias for $\hat{\gamma}_{01}$, and the second panel shows relative percentage bias for $\hat{\gamma}_{10}$. The last panel shows the relative percentage bias in the compositional effect $(\hat{\gamma}_{01} - \hat{\gamma}_{10})$.

First, with respect to measurement model 1 (See Table 5.8), in which the generating values of η_{ij} and ξ_{ij} are analyzed (columns titled G), the bias of $\hat{\gamma}_{01}$ ranged from 1 to 15% across the sampling conditions. The percentage bias is less than 5%, but when the

ICC is small and the number of groups sampled was as small as 25, the bias increases up to about 15%.

As expected, the bias of $\hat{\gamma}_{10}$ is less than 3% across sampling conditions because this parameter is estimated mostly with information from level 1. The two parameter estimates were combined to obtain the compositional effect estimates presented in the last panel. Percentage bias in the compositional effect ranged from about 3 to 50%. When ICC and the number of people per group were small, the bias is about 10%, and it can be up to 50% when ICC and the number of groups sampled were small. Therefore, small ICC conditions are problematic in general. When small ICC is combined with a small number of people per group, the bias gets worse. However, the largest bias occurs when small ICC is combined with a small number of groups, which is to be expected.

Second, the general patterns of relative bias in the parameters of interest from the latent variable modeling approach (see, L titled columns) are similar to those from the generating value analysis. However, the bias is larger in general, as latent factors are estimated by imperfect measures. For measurement model 1, the relative bias of between-level coefficient $\hat{\gamma}_{01}$ can be up to 10% when ICC is small with small or large numbers of people per group. However, when small ICC is combined with a small number of groups, the bias can be as large as 22%. For $\hat{\gamma}_{10}$, the bias is less than 5% across conditions. With respect to the compositional effect, the magnitude of bias is also similar to those from the generating value analysis in that the biggest bias (about 50%) occurs when ICC and the number of groups sampled were small.

Third, the bias when a traditional model is applied (columns titled M), is severe and similar to levels that are reported in previous research. For $\hat{\gamma}_{01}$ and $\hat{\gamma}_{10}$, the percentage bias ranges from 30 to 70%. With respect to the compositional effect, the bias can be small as about 8% when ICC is large and the sampling condition favorable, but the bias can be as large as 80% when the sample is associated with small ICC and a small number of people per group. It is noteworthy that the bias in the compositional effect from the traditional model can be smaller than the bias when a latent variable model is applied to when ICC is small and the number of group sampled is small as 25. However, the number of replication is small and more research is warranted.

Fourth, given that the bias in estimates using generating values is not avoidable, the latent variable modeling approach yields less bias in compositional effect estimates than a traditional analysis. Particulary, when ICC is small and the number of people per group is small, the improved performance of latent variable modeling in estimation of contextual effect is prominent, as long as enough number of groups are sampled. However, in the combination of small ICC and small number of groups, the traditional model may yield less bias then latent variable model approach.

Fifth, performance of the traditional model and the latent variable model in terms of estimating $\hat{\gamma}_{01}$, $\hat{\gamma}_{10}$, and compositional effect, is similar across measurement conditions, indicating the measurement model is a less influential source of bias in this study. Considering the bias of estimates when generating values were analyzed, no significant improvement was found as enforcing predictor measurement model by adding more information (measurement model 2) or items (measurement model 3). Very slight improvement was found with respect to the bias in $\hat{\gamma}_{10}$ from traditional model. However, in terms of latent variable modeling, the bias in the compositional effect when the sample is associated with a small ICC and a small number of groups gets even worse using measurement model 3 (See, the last panel of Table 5.10). This might be caused by the fact that many more parameters are estimated in this model (number of parameter estimated is 84) from $N = 500$ sample. Consequently, the parameter estimates may be less precise.

RMSE

The RMSEs of the compositional effect across conditions and models are summarized in Table 5.11. When generating values are analyzed, the smallest RMSE (0.15) is found when ICC is large with favorable sampling conditions, and the largest (0.69) is found with the combination of small ICC and the small number of groups. As previous reported, RMSEs are small in general when the traditional model is applied, indicating that this model yields consistently biased estimates. The latent variable model analysis resulted in generally large RMSE, ranging from 0.21 (when the ICC is large with favorable sampling conditions) and 2.51 (when the ICC is small and the number of group is small, measurement model 1). Again, measurement structures do not seem to make a significant difference in terms of RMSE, indicating that sampling conditions are the primary cause of bias in this study.

Percentage coverage rate

To examine the performance of standard errors, the 95% CI coverage rate for the true compositional effect was calculated across simulated data conditions and models. Results are summarized in Table 5.12.

When generating values are analyzed, the coverage rates of contextual effect across sample conditions are generally as close to 95%, except for the cases where ICC is small and the number of group sampled is small. In this case the coverage rate can be low as 85%. The coverage rates of latent variable model were also similar to those from generating value analysis, ranging from 88% to 98% for measurement model 1 and 2. When more item parameters need to be estimated, the sample is associated with a small ICC, a and small number of groups are sampled, the coverage rate can be low as about 79%.

Traditional model performance in terms of coverage rate for the contextual effect can be very problematic when the number of people per group and ICC are small at the same time, in that the coverage can be low as 7%. When the number of groups sampled is small and ICC is small, the coverage rate of traditional model was slightly better than latent variable approach for measurement model 3.

Observed percentage of significant compositional effect

To examine how researchers can make different statistical decisions when they apply a traditional model and a latent variable model to different conditions, the percentages of significant compositional effect are calculated. Results for measurement model 1 are shown in Table 5.13.

The first panel of Table 5.13 shows empirical Type I error rates of models across data conditions. Generating value analysis model yields Type I error rates of .05 to .07 across sampling conditions. The latent variable model is similar, except for the cases when the number of people per group is small. When the number of people per group is small and ICC is small, Type I error increases to .14, indicating that it is more likely to conclude that there is a significant contextual effect than other approaches. On the other hand, when the number of people per group is large and ICC is small, latent variable modeling is very conservative, rarely indicating that there is a significant contextual effect.

For traditional model, Type I error rate inflation is huge - up to .57 when ICC is large and the number of people per group is large. Under the conditions when small ICC combines with a small number of group or a small number of people per group, the type I error of the traditional model remains at a proper level. Finally, these trends are consistent across measurement models as summarized in Table 5.14.

When a compositional effect is large (see the third panel), generating value analysis yields power of about .85 when ICC is large and the number of groups is large. When ICC is small the power decreases to as low as .35 with favorable sampling conditions. The lowest power (.15) is found when ICC is small and the number of groups is small.

The patterns are similar for the latent variable model, but when ICC is small, and the number of people per group or the number of groups is small, the latent variable model yields a slightly higher percentage of significant compositional effects. While the traditional model can yield a very high percentage of significant compositional effects when the ICC is large and the number of people per group is large, the power decreases remarkably when ICC is small and the number of people per group or the number of groups is small. Similarly, the power of the traditional model is high when ICC is large and the number of people per group is large, and the power of latent variable model tends to be higher than the traditional model when a small ICC is associated with a small number of people per group.

In summary, relative bias of $\hat{\gamma}_{01}$ are $\hat{\gamma}_{10}$ are large when the traditional model is applied, as found in previous research. However, the difference between the two estimates is kept to a certain degree, since both coefficients are underestimated. As a result, the final compositional effect point estimate can be well captured by 95% confidence intervals when ICC is big and the sampling condition is desirable. However, the Type I error rate is severely inflated under this condition, as this model can falsely claim that there is a significant compositional effect even when there is none. This can be seen as a phenomenon caused by the combination of biased estimates and the spuriously small standard error estimates that the traditional model yields.

On the other hand, bias seems unavoidable when a sampling condition is not favorable and ICC is small, even with generating true scores. However, the latent variable model yields less biased estimates in general. When ICC is small and the number of people per group is small, the Type I error rate slightly increases, but the magnitude is still preferable compared to the Type I error inflation of the traditional model across sampling conditions. The concern with the latent variable model approach in terms of sampling is more about the small number of groups rather than the number of people per group. As long as the number of groups sampled is sufficiently large, the performance of the latent variable model approach is preferable.

Finally, the measurement structure was less influential in this study, and this can be due to quality of all fixed item parameters that are used across simulation studies. However, the results from measurement model 3 in this study indicate that estimation of too many item parameters with limited sample size can possibly undermine the performance of the latent variable model approach.

5.2.2.2 Cross-level Interaction Model

The following results indicate that the nonlinear multilevel latent variable modeling is preferable to the traditional HLM based on the bias, RMSE, coverage rates, and Type I error rates. The point estimates and standard errors are less biased and more precise across sampling conditions when the latent variable model was applied to analyze the simulated data. In particular, when the traditional HLM was applied, bias in the cross-level interaction coefficient was substantial. Moreover, the combination of bias and small standard errors resulted in the failure of most 95% CIs to cover a true generating value.

Relative percentage bias

The relative percentage bias in $\hat{\gamma}_{11}$ across simulated data conditions is summarized in Table 5.15. First, when generating values are analyzed, bias can be as small as about 2% when the sampling condition is favorable and ICC is large enough. However, the bias can be as large as about 40% even when generating values are analyzed when the ICC is small and the number of groups sampled is 25. While the traditional approach yields more than 75% underestimation across conditions and reached almost 100% when a small ICC is combined with limited sample conditions, the bias in $\hat{\gamma}_{11}$ from the latent variable model analysis was smaller than that from the manifest variable model analysis. The smallest bias (about 9%) is found when the ICC is big and the number of groups and the number of people per group are sufficiently large (i.e., 100 and 20, respectively). However, the bias increases to 70% when small ICC is combined with a small number of groups and about 50% when either ICC or sampling condition is not favorable.

RMSE

Table 5.16 contains RMSE of the cross-level interaction across simulated data conditions. RMSEs from the generating value analysis ranged from 0.11 to 0.67. When the latent variable model was applied, RMSEs ranged from 0.19 to 1.05. The largest RMSEs for both models were found when there was a large cross-level interaction but ICC of the latent predictor is small and only 25 groups were sampled from the population. However, with respect to traditional model analysis, RMSEs were large in general when the magnitude of cross-level interaction was big regardless of ICC and sampling condition, indicating that cross-level interaction estimates from the traditional model approach were not only biased but also varied substantially across replications, whereas those from the latent variable model analysis were less biased and relatively consistent across replications.

Percentage coverage rate

Coverage rates for true cross-level interaction effects using 95% confidence intervals are reported in Table 5.17. When generating values were analyzed, 95% confidence intervals covered the true cross-level interaction 81 to 100% of the time. When the latent variable model was applied, the coverage rates ranged from 12 to 87% depending on sampling conditions. When the number of sampled groups was small, the confidence intervals hardly captured the true values, even with the latent variable modeling approach. However, these coverage rates were still much higher than those from the traditional model approach. As bias in estimates was big and the standard error estimates were small in the traditional model approach, it was extremely rare to observe that confidence intervals actually covered the true value. Most of the coverage rates were 0.

Observed percentage of significant cross-level interaction

Table 5.18 shows observed percentage of significant cross-level interaction across different sampling conditions and analysis models. Results from the generating value analyses are encouraging in that power can be about .80 for both large and small cross-level interactions, as long as ICC is large enough and sufficient number of groups is sampled. However, when a small number of groups is sampled, the power can be as low as .32 for a large cross-level interaction and .06 for a small cross-level interaction. The latent variable model approach can detect cross-level interaction better than the traditional modeling approach in that the percentages of significant cross-level interactions are higher in general than those from the traditional model analysis. However, when the cross-level interaction is large and the sampling condition is favorable with large ICC, the traditional model can detect the effect slightly more frequently than the latent variable modeling approach. It should be noted that the CI's do not cover the true value in this case even though the traditional model can detect the existence of the cross-level interaction. It is notable that the power of the traditional model decreases dramatically when either ICC or the number of people per group is small. In terms of empirical Type I error rates, the latent variable model analysis shows the highest Type I error rate (.17) when ICC is small and the number of sampled people per group is small. Except for that condition, Type I error rates for the three models are at acceptable levels.

In summary, the bias in cross-level interaction estimates is smaller and less variable when the latent variable model is applied than when the traditional model is applied. In spite of the variability across sampling conditions, the latent variable modeling approach can better capture the value of true cross-level interaction effect and detect it with significance more often than the traditional model approach. However, when a small number of people per group is sampled and the ICC of the latent predictor is small, the Type I error rate of the latent variable model is inflated; therefore, the analysis can yield a spurious cross-level interaction effect. Traditional model, however, leads to misleading inferences as the CIs do not cover true values.

Table 5.1: Generating values and estimates for a compositional effect model (N=2,000, ng=100, np=20, 10/10 converged)

Structural Parameters							
		EM (5qp)		EM (14qp)		MHRM	
	θ	E($\hat{\theta}$)	E{se($\hat{\theta}$)}	E($\hat{\theta}$)	E{se($\hat{\theta}$)}	E($\hat{\theta}$)	E{se($\hat{\theta}$)}
γ_{01}	1.00	1.02	0.19	1.01	0.19	1.00	0.18
γ_{10}	0.50	0.52	0.05	0.51	0.05	0.52	0.09
τ_{00}	1.00	0.90	0.16	0.91	0.17	0.93	0.16
$var(\xi_{.j})$	0.43	0.40	0.07	0.42	0.07	0.42	0.07
Measurement Parameters							
a_{x1}	0.80	0.79	0.07	0.79	0.07	0.79	0.08
a_{x2}	1.00	1.01	0.08	1.01	0.08	1.00	0.09
a_{x3}	1.20	1.24	0.09	1.24	0.09	1.24	0.11
a_{x4}	1.40	1.39	0.10	1.39	0.10	1.39	0.12
a_{x5}	1.60	1.67	0.14	1.67	0.14	1.69	0.15
a_{y1}	0.80	0.78	0.06	0.78	0.06	0.78	0.06
a_{y2}	1.00	1.00	0.07	1.00	0.07	1.00	0.07
a_{y3}	1.20	1.23	0.09	1.23	0.09	1.23	0.08
a_{y4}	1.40	1.40	0.11	1.40	0.11	1.40	0.10
a_{y5}	1.60	1.61	0.13	1.61	0.13	1.60	0.12
c_{x1}	-0.80	-0.75	0.08	-0.75	0.08	-0.75	0.06
c_{x2}	0.00	0.02	0.08	0.02	0.08	0.02	0.05
c_{x3}	1.20	1.30	0.11	1.30	0.11	1.29	0.08
c_{x4}	-0.70	-0.61	0.11	-0.61	0.11	-0.62	0.07
c_{x5}	0.80	0.92	0.14	0.92	0.14	0.92	0.08
c_{y1}	-0.80	-0.80	0.11	-0.80	0.11	-0.81	0.06
c_{y2}	0.00	0.01	0.13	0.01	0.13	0.00	0.05
c_{y3}	1.20	1.19	0.16	1.19	0.16	1.18	0.08
c_{y4}	-0.70	-0.74	0.18	-0.74	0.18	-0.75	0.07
c_{y5}	0.80	0.79	0.21	0.79	0.21	0.78	0.08
Efficiency							
one processor		5∼7 min		60∼100min		35∼40min	

Note.θ = Generating values; E($\hat{\theta}$) = mean of point estimates; E{se($\hat{\theta}$)} = mean of estimated SEs (post-convergence approximated SEs); a = item slope parameters; c = item threshold parameters.

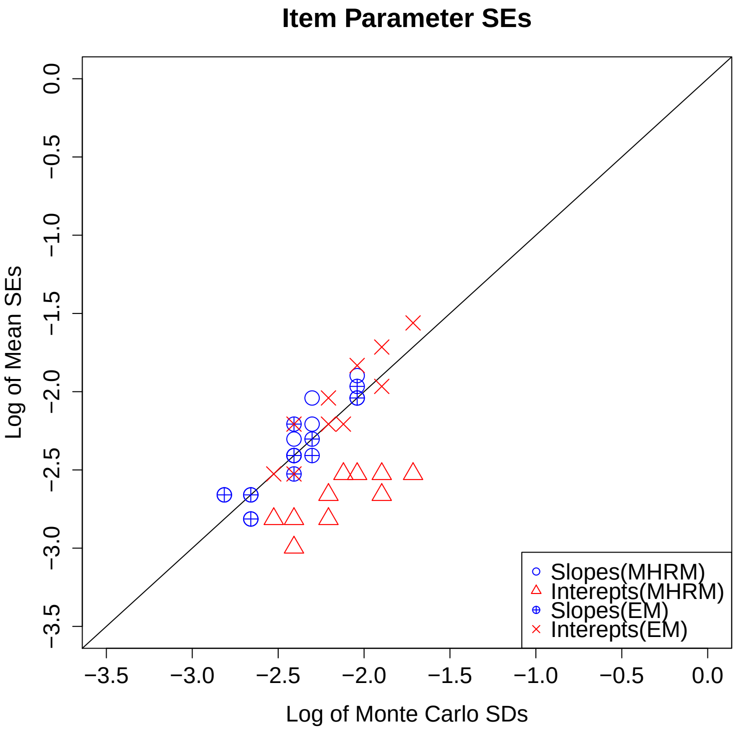


Figure 5.1: Comparisons of standard errors for item parameters

Table 5.2: Generating values and estimates for a compositional effect model (N=2,000, ng=100, np=20)

Structural Parameters						
	θ	E($\hat{\theta}$)	E{se1($\hat{\theta}$)}	SD($\hat{\theta}$)	E{se2($\hat{\theta}$)}	95% Coverage using se1
γ_{01}	1.00	0.99	0.17	0.19	0.18	95.0
γ_{10}	0.50	0.50	0.06	0.07	0.09	95.0
τ_{00}	1.00	0.97	0.20	0.18	0.16	89.0
$var(\xi_{.j})$	0.43	0.43	0.08	0.09	0.07	89.0
Measurement Parameters						
a_{x1}	0.80	0.80	0.07	0.06	0.07	98.0
a_{x2}	1.00	1.01	0.10	0.09	0.09	91.0
a_{x3}	1.20	1.22	0.12	0.10	0.11	92.0
a_{x4}	1.40	1.40	0.12	0.10	0.13	84.0
a_{x5}	1.60	1.60	0.15	0.13	0.15	73.0
a_{y1}	0.80	0.80	0.07	0.07	0.06	95.0
a_{y2}	1.00	1.01	0.07	0.07	0.07	94.0
a_{y3}	1.20	1.21	0.10	0.09	0.09	86.0
a_{y4}	1.40	1.39	0.10	0.09	0.10	89.0
a_{y5}	1.60	1.61	0.10	0.13	0.13	74.0
c_{x1}	0.80	0.80	0.14	0.08	0.06	94.0
c_{x2}	0.00	0.00	0.07	0.09	0.05	95.0
c_{x3}	-1.20	-1.22	0.09	0.12	0.08	91.0
c_{x4}	0.70	0.69	0.12	0.11	0.07	89.0
c_{x5}	-0.80	-0.80	0.12	0.15	0.08	89.0
c_{y1}	0.80	0.81	0.08	0.09	0.06	87.0
c_{y2}	0.00	0.01	0.11	0.11	0.06	78.0
c_{y3}	-1.20	-1.20	0.13	0.13	0.08	75.0
c_{y4}	0.70	0.71	0.15	0.15	0.07	62.0
c_{y5}	-0.80	-0.79	0.14	0.18	0.08	59.0
Efficiency						
			35∼40min		90∼120min	

Note.θ = Generating values; E($\hat{\theta}$) = mean of point estimates; E{se1($\hat{\theta}$)} = mean of recursively approximated standard error estimates (67 converged replications); E{se2($\hat{\theta}$)} = mean of post-convergence approximated standard errors; SD($\hat{\theta}$) = Standard deviation of point estimates; 95% Coverage using se1: Percentage coverage rate of generating value using post-convergence approximated standard errors; a = item slope parameters; c = item threshold parameters.

Table 5.3: Generating values and estimates for a cross-level interaction model (N=2,000, ng=100, np=20, 1st simulated data set)

Structural Parameters							
		EM (5qp)		EM (8qp)		MHRM	
	θ	E($\hat{\theta}$)	E{se($\hat{\theta}$)}	E($\hat{\theta}$)	E{se($\hat{\theta}$)}	E($\hat{\theta}$)	E{se($\hat{\theta}$)}
γ_{01}	1.00	1.86	0.25	1.35	0.22	1.44	0.22
γ_{10}	0.50	1.94	0.15	0.63	0.13	0.63	0.05
γ_{11}	0.50	1.27	0.45	0.83	0.29	0.83	0.06
τ_{00}	1.00	0.85	0.11	0.88	0.12	0.90	0.18
τ_{11}	1.00	0.78	0.33	0.83	0.25	0.79	0.16
τ_{01}	0.50	0.96	0.15	0.49	0.12	0.49	0.11
$var(\xi_{.j})$	0.43	0.40	0.02	0.39	0.05	0.39	0.07
Measurement Parameters							
a_{x1}	0.80	0.78	–	0.78	–	0.78	0.08
a_{x2}	1.00	1.40	0.14	0.96	0.14	0.96	0.07
a_{x3}	1.20	2.05	0.19	1.41	0.19	1.41	0.12
a_{x4}	1.40	2.37	0.21	1.62	0.21	1.63	0.18
a_{x5}	1.60	2.51	0.24	1.69	0.25	1.71	0.12
a_{y1}	0.80	0.79	0.00	0.79	0.00	0.79	0.05
a_{y2}	1.00	0.95	0.11	0.93	0.11	0.93	0.06
a_{y3}	1.20	1.17	0.11	1.15	0.12	1.16	0.07
a_{y4}	1.40	1.00	0.14	0.98	0.15	1.22	0.08
a_{y5}	1.60	1.43	0.18	1.40	0.19	1.51	0.09
c_{x1}	-0.80	-0.68	0.06	-0.73	0.07	-0.74	0.05
c_{x2}	0.00	0.10	0.08	0.10	0.08	0.09	0.05
c_{x3}	1.20	1.43	0.11	1.43	0.12	1.41	0.09
c_{x4}	-0.70	-0.52	0.11	-0.51	0.12	-0.53	0.08
c_{x5}	0.80	1.11	0.13	1.10	0.14	1.09	0.08
c_{y1}	-0.80	-0.72	0.09	-0.73	0.11	-0.73	0.06
c_{y2}	0.00	0.03	0.11	0.04	0.13	0.03	0.06
c_{y3}	1.20	1.26	0.14	1.26	0.16	1.26	0.08
c_{y4}	-0.70	-0.53	0.14	-0.52	0.16	-0.52	0.07
c_{y5}	0.80	0.96	0.17	0.96	0.20	0.96	0.08
Efficiency							
8 processors		15 min		100 min		60min	
1 processor		40 min		4hour 40 min			

Note 1.θ = Generating values; E($\hat{\theta}$) = mean of point estimates; E{se($\hat{\theta}$)} = mean of estimated SEs (post-convergence approximated SEs); a = item slope parameter; c = item threshold parameter.

*Note 2.*Mplus does not allow standardized factor identification option; therefore, anchoring the first factor loading option was used to estimate the model and the results are transformed to make the estimate comparable.

Table 5.4: Generating values and estimates for a cross-level interaction model using MH-RM algorithm (N=2,000, ng=100, np=20, 26/50 converged)

Structural Parameters				
	θ	E($\hat{\theta}$)	E{se($\hat{\theta}$)}	SD($\hat{\theta}$)
γ_{01}	1.00	1.07	0.18	0.21
γ_{10}	0.50	0.55	0.07	0.14
γ_{11}	0.50	0.46	0.27	0.19
τ_{00}	1.00	1.06	0.29	0.17
τ_{11}	1.00	1.05	0.28	0.27
τ_{01}	0.50	0.50	0.15	0.12
$var(\xi_{.j})$	0.43	0.43	0.07	0.09
Measurement Parameters				
a_{x1}	0.80	0.78	0.08	0.06
a_{x2}	1.00	0.98	0.08	0.08
a_{x3}	1.20	1.23	0.11	0.09
a_{x4}	1.40	1.37	0.12	0.14
a_{x5}	1.60	1.59	0.18	0.12
a_{y1}	0.80	0.77	0.06	0.06
a_{y2}	1.00	0.97	0.07	0.06
a_{y3}	1.20	1.19	0.11	0.06
a_{y4}	1.40	1.37	0.12	0.14
a_{y5}	1.60	1.56	0.17	0.13
c_{x1}	-0.80	-0.77	0.06	0.09
c_{x2}	0.00	0.00	0.05	0.09
c_{x3}	1.20	1.21	0.08	0.12
c_{x4}	-0.70	-0.66	0.07	0.14
c_{x5}	0.80	0.78	0.08	0.14
c_{y1}	-0.80	-0.79	0.06	0.12
c_{y2}	0.00	0.00	0.06	0.15
c_{y3}	1.20	1.21	0.09	0.19
c_{y4}	-0.70	-0.67	0.08	0.23
c_{y5}	0.80	0.84	0.09	0.24
Efficiency				
60∼90min				

Note.θ = Generating values; E($\hat{\theta}$) = mean of point estimates; E{se($\hat{\theta}$)} = mean of estimated SEs (post-convergence approximated SEs); a = item slope parameter; c = item threshold parameter.

Table 5.5: Conditions of measurement models

Condition	ξ_{ij} indicators	η_{ij} indicators
Measurement model 1	X1∼X5 (2PL)	Y1∼Y5 (2PL)
Measurement model 2	X1∼X5 (GR, K=5)	Y1∼Y5 (GR, K=5)
Measurement model 3	X1∼X15 (2PL), X16∼X20(GR, K=5)	Y1∼Y5 (GR, K=5)

Table 5.6: Generating values for item parameters

Measurement Model 1		
	slope	intercept
X1, Y1	0.8	-1
X2, Y2	1.0	0
X3, Y3	1.2	1
X4, Y4	1.4	-0.5
X5, Y5	1.6	0.5
Measurement Model 2		
X1, Y1	0.8	-1, 0, 1, 2
X2, Y2	1.0	-1, 0, 1, 2
X3, Y3	1.2	-1, 0, 1, 2
X4, Y4	1.4	-1, 0, 1, 2
X5, Y5	1.6	-1, 0, 1, 2
Measurement Model 3		
X1	0.6	-1
X2	0.8	-0.5
X3	1.0	0
X4	1.2	0.5
X5	1.4	1
X6	1.6	2
X7	1.8	-1
X8	0.6	-0.5
X9	0.8	0
X10	1.0	0.5
X11	1.2	1
X12	1.4	2
X13	1.6	0.5
X14	1.8	1
X15	1.5	2
X16, Y1	0.8	-1, 0, 1, 2
X17, Y2	1.0	-1, 0, 1, 2
X18, Y3	1.2	-1, 0, 1, 2
X19, Y4	1.4	-1, 0, 1, 2
X20, Y5	1.6	-1, 0, 1, 2

Table 5.7: Percentage of convergence and average time per estimation (sec)

Compositional effect model						
Large Compositional Effect = 0.5						
	np=20			np=5		
ng=100	M1	M2	M3	M1	M2	M3
ICC=0.1	100(2781)	89(4911)	96(24000)	97(972)	81(1593)	91(5934)
ICC=0.3	100(2657)	95(5301)	93(24450)	100(955)	95(1613)	94(6022)
ng=25	M1	M2	M3	M1	M2	M3
ICC=0.1	98(1046)	92(1522)	97(6253)		N/A	
ICC=0.3	99(865)	93(1524)	98(6407)			
Small Compositional Effect = 0.2						
	np=20			np=5		
ng=100	M1	M2	M3	M1	M2	M3
ICC=0.1	97(2937)	91(5165)	94(21730)	95(1021)	92(1588)	90(4909)
ICC=0.3	98(1785)	92(4910)	99(22060)	100(1046)	91(1593)	95(4910)
ng=25	M1	M2	M3	M1	M2	M3
ICC=0.1	95(919)	78(1521)	92(5922)		N/A	
ICC=0.3	93(915)	95(1519)	95(5219)			
Cross-level interaction model						
Large Cross-level interaction = 1						
	np=20			np=5		
ng=100	M1	M2	M3	M1	M2	M3
ICC=0.1	56 (3412)	30 (28210)	4 (86720)	76 (1674)	40 (8654)	16 (25810)
ICC=0.3	48 (2426)	0 (75789)	16 (28860)	45 (1956)	32 (9357)	16 (27850)
ng=25	M1	M2	M3	M1	M2	M3
ICC=0.1	52(1639)	42 (7247)	15 (27590)		N/A	
ICC=0.3	58(2103)	32 (10630)	16 (29190)			
Small Cross-level interaction = 0.5						
	np=20			np=5		
ng=100	M1	M2	M3	M1	M2	M3
ICC=0.1	22 (2370)	45 (22120)	4 (93150)	51 (1282)	32 (5901)	15 (27270)
ICC=0.3	33 (2549)	17 (20460)	5 (80010)	33 (1604)	18 (6769)	15 (26850)
ng=25	M1	M2	M3	M1	M2	M3
ICC=0.1	38 (962)	28 (5880)	14 (28750)		N/A	
ICC=0.3	35 (1684)	35 (8753)	14 (28960)			

Note1. M1 = Measurement model 1 result; M2 = Measurement model 2 result; M3 = Measurement model 3 result; ng = number of groups; np = number of people per group.

Note2. 100 replications for the compositional effect model and 50 replications for the cross-level interaction model.

Note3. For cross-level interaction model with M3 condition, any of replication didn't converge in terms of standard errors. The reported numbers are merely the number of replications that have been made.

Table 5.8: Relative percentage bias of $\hat{\gamma}_{01}$, $\hat{\gamma}_{10}$, and compositional effect ($\hat{\gamma}_{01} - \hat{\gamma}_{10}$), measurement model 1

	$\gamma_{01} = 1$											
	Large Compositional Effect = 0.5						Small Compositional Effect = 0.2					
	np=20			np=5			np=20			np=5		
ng=100	G	L	M	G	L	M	G	L	M	G	L	M
ICC=0.1	-1.41	9.95	-47.05	-1.46	-0.31	-68.20	-1.77	2.51	-47.64	1.75	-1.67	-60.09
ICC=0.3	-1.46	-1.06	-32.91	0.89	14.04	-47.09	-0.31	1.25	-34.43	-0.64	5.26	-45.60
ng=25	G	L	M	G	L	M	G	L	M	G	L	M
ICC=0.1	14.17	21.77	-42.07		N/A		-10.55	6.33	-58.52		N/A	
ICC=0.3	0.05	3.27	-32.46				-3.65	-1.24	-35.45			
	$\gamma_{10} = 0.5$						$\gamma_{10} = 0.8$					
	Large Compositional Effect = 0.5						Small Compositional Effect = 0.2					
	np=20			np=5			np=20			np=5		
ng=100	G	L	M	G	L	M	G	L	M	G	L	M
ICC=0.1	0.40	0.86	-58.41	-0.15	1.79	-59.06	0.04	-0.08	-61.29	0.09	2.55	-60.65
ICC=0.3	-0.22	0.20	-58.32	0.18	2.15	-58.84	0.27	1.66	-60.28	0.38	0.88	-60.20
ng=25	G	L	M	G	L	M	G	L	M	G	L	M
ICC=0.1	0.50	2.69	-58.29		N/A		-1.11	-5.00	-62.33		N/A	
ICC=0.3	0.06	-1.02	-58.70				-0.37	-0.15	-60.14			
	Compositional Effect (γ_{01} - γ_{10})											
	Large Compositional Effect = 0.5						Small Compositional Effect = 0.2					
	np=20			np=5			np=20			np=5		
ng=100	G	L	M	G	L	M	G	L	M	G	L	M
ICC=0.1	-3.22	19.04	-35.68	-2.76	-2.42	-77.35	-9.03	12.86	6.93	8.38	-18.56	-57.85
ICC=0.3	-2.71	-2.32	-7.49	1.60	25.92	-35.34	-2.60	-0.38	68.99	-4.68	22.75	12.82
ng=25	G	L	M	G	L	M	G	L	M	G	L	M
ICC=0.1	27.83	40.00	-25.86		N/A		-48.31	50.77	-18.02		N/A	
ICC=0.3	0.04	7.57	-6.21				-16.78	-5.61	63.31			

Note. G = Generating value analysis result; L = Latent variable model analysis result; M = Manifest variable model analysis result; ng = number of groups; np = number of people per group.

Table 5.9: Relative percentage bias of $\hat{\gamma}_{01}$, $\hat{\gamma}_{10}$, and compositional effect ($\hat{\gamma}_{01} - \hat{\gamma}_{10}$), measurement model 2

						$\gamma_{01} = 1$						
	Large Compositional Effect = 0.5						Small Compositional Effect = 0.2					
	np=20			np=5			np=20			np=5		
ng=100	G	L	M	G	L	M	G	L	M	G	L	M
ICC=0.1	-1.28	3.14	-47.98	-3.25	2.23	-62.25	3.91	7.36	-44.14	-3.57	2.35	-54.62
ICC=0.3	-1.94	-1.06	-32.67	-1.68	3.61	-45.13	1.04	1.59	-33.49	-0.30	1.81	-42.37
ng=25	G	L	M	G	L	M	G	L	M	G	L	M
ICC=0.1	-9.06	8.62	-54.05		N/A		-15.32	-22.02	-55.50		N/A	
ICC=0.3	2.44	7.89	-30.69				-2.90	1.93	-35.78			
	$\gamma_{10} = 0.5$						$\gamma_{10} = 0.8$					
	Large Compositional Effect = 0.5						Small Compositional Effect = 0.2					
	np=20			np=5			np=20			np=5		
ng=100	G	L	M	G	L	M	G	L	M	G	L	M
ICC=0.1	0.87	0.90	-51.87	1.68	0.75	-50.54	-0.15	-1.40	-55.56	-0.36	-0.40	-54.82
ICC=0.3	-0.24	-0.68	-50.55	1.00	0.30	-50.10	-0.15	-0.60	-53.96	-0.04	-0.28	-51.97
ng=25	G	L	M	G	L	M	G	L	M	G	L	M
ICC=0.1	-2.80	-2.67	-52.09		N/A		0.34	-1.58	-54.43		N/A	
ICC=0.3	-0.97	-1.33	-51.81				0.31	-0.54	-52.52			
						Compositional Effect (γ_{01} - γ_{10})						
	Large Compositional Effect = 0.5						Small Compositional Effect = 0.2					
	np=20			np=5			np=20			np=5		
ng=100	G	L	M	G	L	M	G	L	M	G	L	M
ICC=0.1	-3.44	5.37	-44.09	-8.18	3.71	-73.95	20.14	42.42	1.55	-16.44	13.35	-53.86
ICC=0.3	-3.63	-1.44	-14.80	-4.36	6.92	-40.16	5.80	10.34	48.37	-1.34	10.17	-3.93
ng=25	G	L	M	G	L	M	G	L	M	G	L	M
ICC=0.1	-15.33	19.90	-56.00		N/A		-77.92	-103.78	-59.79		N/A	
ICC=0.3	5.85	17.11	-9.57				-15.75	11.79	31.20			

Note. G = Generating value analysis result; L = Latent variable model analysis result; M = Manifest variable model analysis result; ng = number of groups; np = number of people per group.

Table 5.10: Relative percentage bias of $\hat{\gamma}_{01}$, $\hat{\gamma}_{10}$, and compositional effect ($\hat{\gamma}_{01} - \hat{\gamma}_{10}$), measurement model 3

	$\gamma_{01} = 1$											
	Large Compositional Effect = 0.5						Small Compositional Effect = 0.2					
	np=20			np=5			np=20			np=5		
ng=100	G	L	M	G	L	M	G	L	M	G	L	M
ICC=0.1	1.95	10.17	-53.85	0.38	10.91	-66.73	3.74	4.24	-51.52	4.15	9.24	-58.64
ICC=0.3	1.42	1.40	-48.02	-0.56	4.23	-56.56	-0.04	0.39	-46.71	2.82	3.84	-52.33
ng=25	G	L	M	G	L	M	G	L	M	G	L	M
ICC=0.1	14.84	64.97	-48.66		N/A		2.31	-20.40	-53.33		N/A	
ICC=0.3	-3.05	-0.23	-49.17				3.27	6.39	-46.15			
	$\gamma_{10} = 0.5$						$\gamma_{10} = 0.8$					
	Large Compositional Effect = 0.5						Small Compositional Effect = 0.2					
	np=20			np=5			np=20			np=5		
ng=100	G	L	M	G	L	M	G	L	M	G	L	M
ICC=0.1	-0.09	0.78	-54.94	-1.02	-1.71	-55.46	-0.31	-0.05	-56.07	-0.97	-1.86	-56.25
ICC=0.3	-0.36	2.23	-55.20	-0.28	0.21	-55.44	0.08	0.17	-56.37	-0.26	-1.41	-57.06
ng=25	G	L	M	G	L	M	G	L	M	G	L	M
ICC=0.1	-1.41	-2.16	-56.60		N/A		-0.76	-1.60	-56.96		N/A	
ICC=0.3	1.57	1.64	-54.86				0.91	2.61	-55.65			
	Compositional Effect (γ_{01} - γ_{10})											
	Large Compositional Effect = 0.5						Small Compositional Effect = 0.2					
	np=20			np=5			np=20			np=5		
ng=100	G	L	M	G	L	M	G	L	M	G	L	M
ICC=0.1	3.99	19.56	-52.77	1.78	23.52	-77.99	19.97	21.41	-33.31	24.65	53.63	-68.18
ICC=0.3	3.19	0.57	-40.84	-0.85	8.26	-57.69	-0.52	1.30	-8.05	15.10	24.83	-33.41
ng=25	G	L	M	G	L	M	G	L	M	G	L	M
ICC=0.1	31.09	132.09	-40.71		N/A		14.59	-95.60	-38.82		N/A	
ICC=0.3	-7.68	-2.10	-43.48				12.73	21.50	-8.15			

Note. G = Generating value analysis result; L = Latent variable model analysis result; M = Manifest variable model analysis result; ng = number of groups; np = number of people per group.

Table 5.11: Root-mean-square errors of compositional effect ($\hat{\gamma}_{01} - \hat{\gamma}_{10}$)

	Measurement model 1											
	Large Compositional Effect = 0.5						Small Compositional Effect = 0.2					
	np=20			np=5			np=20			np=5		
ng=100	G	L	M	G	L	M	G	L	M	G	L	M
ICC=0.1	0.28	0.77	0.24	0.31	1.15	0.41	0.32	0.53	0.18	0.34	0.87	0.17
ICC=0.3	0.16	0.21	0.12	0.19	0.51	0.21	0.15	0.30	0.17	0.17	0.40	0.10
ng=25	G	L	M	G	L	M	G	L	M	G	L	M
ICC=0.1	0.63	2.51	0.35		N/A		0.68	1.69	0.37		N/A	
ICC=0.3	0.29	0.43	0.22				0.33	0.46	0.25			
	Measurement model 2											
ng=100	G	L	M	G	L	M	G	L	M	G	L	M
ICC=0.1	0.33	0.66	0.29	0.39	0.88	0.40	0.28	0.56	0.18	0.40	0.80	0.18
ICC=0.3	0.15	0.21	0.13	0.13	0.30	0.22	0.17	0.28	0.15	0.17	0.38	0.11
ng=25	G	L	M	G	L	M	G	L	M	G	L	M
ICC=0.1	0.69	1.54	0.48		N/A		0.60	1.57	0.36		N/A	
ICC=0.3	0.31	0.42	0.22				0.34	0.46	0.22			
	Measurement model 3											
ng=100	G	L	M	G	L	M	G	L	M	G	L	M
ICC=0.1	0.31	0.64	0.29	0.35	0.92	0.41	0.34	0.48	0.16	0.37	0.78	0.18
ICC=0.3	0.20	0.22	0.22	0.20	0.35	0.31	0.17	0.29	0.09	0.20	0.35	0.12
ng=25	G	L	M	G	L	M	G	L	M	G	L	M
ICC=0.1	0.55	1.54	0.30		N/A		0.66	1.62	0.31		N/A	
ICC=0.3	0.39	0.44	0.30				0.31	0.39	0.15			

Note. G = Generating value analysis result; L = Latent variable model analysis result; M = Manifest variable model analysis result; ng = number of groups; np = number of people per group.

Table 5.12: Percentage coverage rate for compositional effect ($\hat{\gamma}_{01} - \hat{\gamma}_{10}$)

	Measurement model 1											
	Large Compositional Effect = 0.5						Small Compositional Effect = 0.2					
	np=20			np=5			np=20			np=5		
ng=100	G	L	M	G	L	M	G	L	M	G	L	M
ICC=0.1	96.0	96.0	91.0	94.8	76.3	30.9	95.9	95.9	95.9	93.9	82.9	92.7
ICC=0.3	98.0	98.0	99.0	94.0	86.0	74.0	96.9	98.0	85.7	94.0	89.0	98.0
ng=25	G	L	M	G	L	M	G	L	M	G	L	M
ICC=0.1	95.8	85.9	100.0		N/A		84.4	93.8	96.9		N/A	
ICC=0.3	96.0	99.0	98.0				94.6	97.8	93.5			
	Measurement model 2											
	Large Compositional Effect = 0.5						Small Compositional Effect = 0.2					
	np=20			np=5			np=20			np=5		
ng=100	G	L	M	G	L	M	G	L	M	G	L	M
ICC=0.1	93.3	92.1	95.5	92.6	87.7	49.4	97.8	94.5	98.9	90.1	88.9	93.8
ICC=0.3	93.7	100.0	97.9	100.0	93.7	72.6	95.7	96.7	91.3	95.6	91.2	97.8
ng=25	G	L	M	G	L	M	G	L	M	G	L	M
ICC=0.1	87.0	92.6	92.6		N/A		96.2	88.5	98.7		N/A	
ICC=0.3	97.8	97.8	97.8				94.7	95.8	100.0			
	Measurement model 3											
	Large Compositional Effect = 0.5						Small Compositional Effect = 0.2					
	np=20			np=5			np=20			np=5		
ng=100	G	L	M	G	L	M	G	L	M	G	L	M
ICC=0.1	93.1	86.2	44.8	92.3	73.6	6.6	94.2	80.8	82.7	92.2	72.2	67.8
ICC=0.3	82.8	93.1	27.6	91.5	73.4	10.6	90.4	88.5	86.5	94.7	74.7	78.9
ng=25	G	L	M	G	L	M	G	L	M	G	L	M
ICC=0.1	95.7	78.7	87.2		N/A		93.8	80.0	90.8		N/A	
ICC=0.3	88.8	85.7	74.5				98.9	94.7	96.8			

Note. G = Generating value analysis result; L = Latent variable model analysis result; M = Manifest variable model analysis result; ng = number of groups; np = number of people per group.

Table 5.13: Percentage of statistically significant compositional effects, measurement model 1

No Compositional Effect = 0						
	np=20			np=5		
ng=100	G	L	M	G	L	M
ICC=0.1	2.9	2.9	21.1	7.0	14.0	1.0
ICC=0.3	5.0	2.0	57.0	5.1	0.0	17.2
ng=25	G	L	M	G	L	M
ICC=0.1	3.0	5.0	5.0		N/A	
ICC=0.3	6.0	4.0	32.0			
Small Compositional Effect = 0.2						
	np=20			np=5		
ng=100	G	L	M	G	L	M
ICC=0.1	9.3	5.2	16.5	7.3	15.9	3.7
ICC=0.3	18.4	13.3	81.6	17.0	22.0	43.0
ng=25	G	L	M	G	L	M
ICC=0.1	12.5	18.8	12.5		N/A	
ICC=0.3	10.8	4.3	24.7			
Large Compositional Effect = 0.5						
	np=20			np=5		
ng=100	G	L	M	G	L	M
ICC=0.1	35.0	28.0	34.0	29.9	34.0	11.3
ICC=0.3	85.0	64.0	96.0	82.0	55.0	71.0
ng=25	G	L	M	G	L	M
ICC=0.1	15.5	26.8	9.9		N/A	
ICC=0.3	33.3	25.3	40.4			

Note. G = Generating value analysis result; L = Latent variable model analysis result; M = Manifest variable model analysis result; ng = number of groups; np = number of people per group.

Table 5.14: Percentage of statistically significant compositional effects, measurement model 2 and 3

	Measurement model 2											
	Large Compositional Effect = 0.5						Small Compositional Effect = 0.2					
	np=20			np=5			np=20			np=5		
ng=100	G	L	M	G	L	M	G	L	M	G	L	M
ICC=0.1	33.71	28.09	23.60	30.86	30.86	7.41	9.89	20.88	12.09	8.64	25.93	6.17
ICC=0.3	84.21	63.16	90.53	80.00	57.89	61.05	27.17	30.43	53.26	23.08	20.88	27.47
ng=25	G	L	M	G	L	M	G	L	M	G	L	M
ICC=0.1	12.96	25.93	11.11		N/A		1.28	14.10	1.28		N/A	
ICC=0.3	33.33	36.56	33.33				9.47	13.68	9.47			
	Measurement model 3											
	Large Compositional Effect = 0.5						Small Compositional Effect = 0.2					
	np=20			np=5			np=20			np=5		
ng=100	G	L	M	G	L	M	G	L	M	G	L	M
ICC=0.1	41.4	48.3	44.8	33.0	38.5	20.9	11.5	25.0	21.2	11.1	21.1	8.9
ICC=0.3	82.8	82.8	93.1	75.5	63.8	68.1	23.1	21.2	65.4	30.5	31.6	40.0
ng=25	G	L	M	G	L	M	G	L	M	G	L	M
ICC=0.1	14.9	31.9	21.3		N/A		9.2	18.5	12.3		N/A	
ICC=0.3	35.7	34.7	44.9				12.6	14.7	24.2			

Note. G = Generating value analysis result; L = Latent variable model analysis result; M = Manifest variable model analysis result; ng = number of groups; np = number of people per group.

Table 5.15: Relative percentage bias of cross-level interaction effect ($\hat{\gamma}_{11}$)

Measurement model 1												
	Small γ_{11} = 0.5						Large γ_{11} = 1					
	np=20			np=5			np=20			np=5		
ng=100	G	L	M	G	L	M	G	L	M	G	L	M
ICC=0.1	-5.82	-45.52	-92.08	6.84	-38.85	-95.77	-14.95	-36.13	-89.85	-1.03	-44.30	-95.26
ICC=0.3	-4.97	-8.52	-76.54	-7.86	-29.35	-86.84	-3.05	-8.35	-72.27	-2.78	-31.19	-86.25
ng=25	G	L	M	G	L	M	G	L	M	G	L	M
ICC=0.1	-41.03	-70.55	-98.45		N/A		-13.37	-24.48	-85.72		N/A	
ICC=0.3	-1.49	-16.55	-78.04				-9.84	-9.13	-76.30			
Measurement model 2												
	Small γ_{11} = 0.5						Large γ_{11} = 1					
	np=20			np=5			np=20			np=5		
ng=100	G	L	M	G	L	M	G	L	M	G	L	M
ICC=0.1	-70.85	-126.10	-100.70	12.96	26.78	-98.44	-5.58	-17.63	-94.17	-6.56	10.66	-95.81
ICC=0.3	0.84	-13.92	-88.82	-3.25	21.55	-92.31	-1.97	-4.58	-89.77	1.25	9.47	-93.79
ng=25	G	L	M	G	L	M	G	L	M	G	L	M
ICC=0.1	-62.94	23.33	-97.14		N/A		8.65	-2.46	-93.47		N/A	
ICC=0.3	3.53	51.19	-89.53				-12.20	-12.37	-90.67			
Measurement model 3												
	Small γ_{11} = 0.5						Large γ_{11} = 1					
	np=20			np=5			np=20			np=5		
ng=100	G	L	M	G	L	M	G	L	M	G	L	M
ICC=0.1	63.26	-57.94	-96.23	-24.98	-1.23	-98.52	21.93	22.89	-96.43	-10.56	-26.19	-98.76
ICC=0.3	-13.77	4.82	-96.97	-13.77	4.82	-96.97	-5.26	-18.84	-95.58	6.37	0.10	-96.68
ng=25	G	L	M	G	L	M	G	L	M	G	L	M
ICC=0.1	-66.02	-58.97	-99.19		N/A		27.75	24.20	-96.02		N/A	
ICC=0.3	-4.99	-28.83	-95.78				3.41	-27.03	-95.39			

Note. G = Generating value analysis result; L = Latent variable model analysis result; M = Manifest variable model analysis result; ng = number of groups; np = number of people per group.

Table 5.16: Root-mean-square errors of cross-level interaction effect ($\hat{\gamma}_{11}$)

Measurement model 1												
	Small γ_{11} = 0.5						Large γ_{11} = 1					
	np=20			np=5			np=20			np=5		
ng=100	G	L	M	G	L	M	G	L	M	G	L	M
ICC=0.1	0.37	0.45	0.96	0.35	0.57	0.98	0.32	0.38	0.90	0.31	0.49	0.96
ICC=0.3	0.16	0.20	0.88	0.16	0.29	0.94	0.12	0.44	0.73	0.21	0.46	0.87
ng=25	G	L	M	G	L	M	G	L	M	G	L	M
ICC=0.1	0.62	0.60	1.00		N/A		0.67	1.05	0.89		N/A	
ICC=0.3	0.35	0.41	0.90				0.39	0.65	0.77			
Measurement model 2												
	Small γ_{11} = 0.5						Large γ_{11} = 1					
	np=20			np=5			np=20			np=5		
ng=100	G	L	M	G	L	M	G	L	M	G	L	M
ICC=0.1	0.49	0.71	1.00	0.34	0.81	0.99	0.27	0.45	0.94	0.38	0.92	0.96
ICC=0.3	0.37	0.22	0.95	0.16	0.41	0.96	0.09	0.51	0.90	0.20	0.72	0.94
ng=25	G	L	M	G	L	M	G	L	M	G	L	M
ICC=0.1	0.71	1.01	0.99		N/A		0.56	0.87	0.94		N/A	
ICC=0.3	0.31	0.59	0.95				0.33	0.47	0.91			
Measurement model 3												
	Small γ_{11} = 0.5						Large γ_{11} = 1					
	np=20			np=5			np=20			np=5		
ng=100	G	L	M	G	L	M	G	L	M	G	L	M
ICC=0.1	0.42	0.60	0.98	0.47	0.97	0.99	0.33	0.78	0.96	0.45	0.77	0.99
ICC=0.3	0.17	0.23	0.98	0.17	0.23	0.98	0.08	0.34	0.96	0.22	0.57	0.97
ng=25	G	L	M	G	L	M	G	L	M	G	L	M
ICC=0.1	0.62	1.02	1.00		N/A		0.67	1.32	0.96		N/A	
ICC=0.3	0.23	0.35	0.98				0.35	0.50	0.95			

Note. G = Generating value analysis result; L = Latent variable model analysis result; M = Manifest variable model analysis result; ng = number of groups; np = number of people per group.

Table 5.17: Percentage coverage rates for cross-level interaction effect ($\hat{\gamma}_{11}$)

Measurement model 1												
	Small γ_{11} = 0.5						Large γ_{11} = 1					
	np=20			np=5			np=20			np=5		
ng=100	G	L	M	G	L	M	G	L	M	G	L	M
ICC=0.1	92.6	59.3	0.0	100.0	65.8	0.0	100.0	30.0	0.0	100.0	51.0	0.0
ICC=0.3	100.0	76.5	0.0	94.7	81.6	0.0	100.0	86.7	0.0	88.9	55.6	0.0
ng=25	G	L	M	G	L	M	G	L	M	G	L	M
ICC=0.1	92.2	19.6	0.0		N/A		89.5	65.8	2.6		N/A	
ICC=0.3	81.0	12.1	0.0				90.6	75.0	0.0			

Note. G = Generating value analysis result; L = Latent variable model analysis result; M = Manifest variable model analysis result; ng = number of groups; np = number of people per group.

Table 5.18: Percentage of significant cross-level interaction effect ($\hat{\gamma}_{11}$), measurement model 1

No cross-level interaction = 0						
	np=20			np=5		
ng=100	G	L	M	G	L	M
ICC=0.1	0.0	7.0	0.0	1.0	17.0	1.0
ICC=0.3	3.0	9.0	1.0	2.0	5.0	0.0
ng=25	G	L	M	G	L	M
ICC=0.1	3.0	3.0	0.0		N/A	
ICC=0.3	0.0	2.0	0.0			
Small cross-level interaction = 0.5						
	np=20			np=5		
ng=100	G	L	M	G	L	M
ICC=0.1	40.7	22.2	7.4	34.2	42.1	3.9
ICC=0.3	82.4	70.6	47.1	84.2	42.1	13.2
ng=25	G	L	M	G	L	M
ICC=0.1	5.9	9.8	2.0		N/A	
ICC=0.3	25.9	22.4	13.8			
Large cross-level interaction = 1						
	np=20			np=5		
ng=100	G	L	M	G	L	M
ICC=0.1	81.8	72.7	0.0	80.4	60.8	3.9
ICC=0.3	100.0	81.3	100.0	100.0	59.3	48.1
ng=25	G	L	M	G	L	M
ICC=0.1	31.6	31.6	18.4		N/A	
ICC=0.3	78.1	46.9	53.1			

Note. G = Generating value analysis result; L = Latent variable model analysis result; M = Manifest variable model analysis result; ng = number of groups; np = number of people per group.

CHAPTER 6

Empirical Applications

To illustrate how nonlinear multilevel latent variable modeling can be applied to empirical data to estimate a contextual effect not only as a compositional effect but also as a cross-level interaction, the following real data analyses were conducted. This chapter summarizes the methods and results of two empirical applications.

6.1 Compositional Effect Model: A "Big-fish-little-pond" Effect

A compositional effect in the relationship between academic self-concept and academic achievement has attracted Marsh et al. (2009)'s attention and been studied in light of multilevel latent variable modeling to address methodological issues. However, a nonlinear measurement structure with a number of categorical indicators has not been studied in the estimation of the compositional effect. In contrast to the previous research of Marsh et al. (2009), in which three continuous indicators were used to measure academic self-concept and academic achievement, the analysis presented here uses categorical item response data for both latent variables, *academic self-concept* and *academic achievement*.

6.1.1 Data

For this compositional effect analysis, a subset of The Programme for International Student Assessment (PISA 2000; OECD, 2000) data were extracted and analyzed. PISA is a large international educational survey. The focus of PISA 2000 was literacy. A large amount of student and school level information that covering cognitive and affective domains was collected.

Though 42 countries participated in the data collection, a sample of students from the United States was analyzed in this study for the purpose of illustration. Originally, a total of 129 reading items were administered to estimate country level *reading literacy* mean using a balanced incomplete block design. However, for simplicity, only booklets 8 and 9 were used for this analysis. These two booklets included 33 reading items, but 1 item was dropped, since all item responses for this item were zero, which meant the item had no information. Therefore, the analyzed data set contained 32 item responses (3 graded responses items with 3 categories and 29 dichotomously scored items) of 667 students from 141 schools. The number of students within a school ranged from 1 to 8, which is rather a small number of students per group. The outcome variable *self concept in reading* was measured by the following three items:

CC02Q05 "I'm hopeless in <test language> classes" (reverse coded),

CC02Q09 "I learn things quickly in <test language> class",

CC02Q23 "I get good marks in <test language>", and

Each item has a Likert-type scale, ranging from 1 (disagree) to 4 (agree). <test language> was English for students in the United States.

6.1.2 Results

The structural parameter estimates from the multilevel latent variable model analysis (EM algorithm and the MH-RM algorithm) and traditional multilevel model analysis are reported in Table 6.1. In general, a positive and significant within-level coefficient $\hat{\gamma}_{10}$ is found across different models and algorithms. Between-level coefficient $\hat{\gamma}_{01}$ estimates were not significantly different from 0 when the multilevel latent model was applied, while the estimate was significantly different from 0 when the traditional multilevel was applied, due to the small standard error.

The compositional effect ("big-fish-little-pond") is calculated by subtracting $\hat{\gamma}_{10}$ from $\hat{\gamma}_{01}$ as illustrated in Figure 6.1. The direction of the compositional was negative as reported in previous research (Marsh et al., 2009). This indicates that two students who have the same levels of achievement can have different level of academic self-concept, depending on the group-level academic achievement. As the compositional effect is negative, the students who belong to a higher-level achievement group tend to have lower academic self-concept compared to students who belong to a lower-level achievement group. On the other hand, the students who belong to a lower-level achievement group tend to have higher academic self-concept compared to students who belong to a higher-level achievement group - just like a fish that feels big if the pond where it lives is small. However, in terms of the statistical significance of the compositional effect, the traditional model yields that the effect is not significantly different from 0. This result is consistent with what was found in the simulation study presented in Chapter 5 in that the power of the latent variable model to detect a compositional effect is higher than that of the traditional model, when the data set is associated with a sufficiently large number of groups and a small number of students per group.

The measurement parameter estimates are summarized in Table 6.2. The point estimates from the MH-RM algorithm are plotted against those from the EM algorithm in Figure 6.2. As can be seen, the estimates are very close to each other. Standard errors of the item parameters exhibited a similar pattern as found previously (see Figure 6.3), confirming that the post-convergence approximation method yields slightly smaller standard errors, while the recursive approximation tends to yield larger standard errors.

6.2 Cross-level Interaction Model: Co-operative Learning Preference and Reading Literacy

As illustrated in Section 1.1.3, a group-level latent variable can influence not only individual-level outcome, but also the relationship between a predictor and an outcome. In particular, when the relation between two variables varies across groups and can be explained as a function of a group-level predictor, a cross-level interaction model is useful to explain the phenomenon. For illustrative purposes, the relationship between *cooperative learning preference* and *academic achievement was examined* was examined. The co-operative learning theory is based upon substantial theoretical and practical foundations in education. An extensive review of the relevant theoretical review is not included in this study, since the substantive implications of the research were not of primary interest of the current study. The effect of co-operative learning has been studied often in the context of higher- or lower-achievement groups (e.g., Tsay & Brady, 2010; Johnson & Johnson, 1989; Baker & Clark, 2010). The application study being reported here is rooted in the awareness of this research stream. However, it should be also noted that co-operative learning preference in this study is different from co-operative learning instruction that is executed in classrooms. The concept of co-operative learning preference is developed in light of learning style or preference in learning situations rather than class instruction. More information related to measures and concept can be found in Owens and Barnes (1992).

6.2.1 Data

For this cross-level interaction model analysis, a subset of PISA 2000 was extracted and analyzed. The data were collected in Korea, and those students who were administered booklets 8 and 9 for reading literacy were used in this analysis. In the process of data cleaning, 4 reading items were dropped, since all item responses were zero. 29 item responses (3 graded responses and 26 dichotomously scored items) of 1,103 students in 143 schools were analyzed. These 29 items are the indicators for the latent predictor variable. The number of students within a school ranged from 1 to 8, which can be considered a small number of students per group. The outcome variable, *co-operative learning preference*, was measured by the following four items:

CC02Q02 “I like to work with other students”,

CC02Q08 “I learn the most when I work with other students”,

CC02Q19 “I like to help other people do well in a group”,

CC02Q19 “It is helpful to put together everyone’s ideas when working on a project”.

Each item has a Likert-type scale, ranging from 1 (disagree) to 4 (agree).

6.2.2 Results

The structural parameter estimates from the multilevel latent variable model analysis (EM algorithm and the MH-RM algorithm) and traditional multilevel model analysis are reported in Table 6.3. In general, positive within- and between-level coefficients ($\hat{\gamma}_{10}$ and $\hat{\gamma}_{01}$) were found, indicating that the level of co-operative learning preference and reading literacy is positively associated. However, none of these were statistically significant when the MH-RM algorithm was applied, and only the between-level coefficient was significant at $p < .05$ level when the EM algorithm was applied, which is also different from the traditional HLM analysis in that both coefficients are statistically different from 0 due to the small standard errors.

The parameter estimate of interest that captures a cross-level interaction effect was $\hat{\gamma}_{11}$, which appears to be negative in this particular example across computational algorithms and models. The negative cross-level interaction can be interpreted as that the relationship between co-operative learning preference and reading literacy is weaker in schools with higher achievement levels, indicating the slope of between two variables becomes less stiffer as school-level achievement increases (see Figure 6.4). If the negative cross-level interaction size is large enough, the direction of the relationship between the co-cooperative learning preference and reading literacy could be negative at schools where school-level reading literacy is very high. However, $\hat{\gamma}_{11}$ was not statistically different from 0 across models and computational algorithms.

With respect to computation, 8 adaptive quadrature points estimation using Mplus did not converge, and only 5-quadrature-point solution was available with some changes in default settings that are related to the M-step. When the MH-RM algorithm was applied, it took 18 hours to estimate, and a large number of samples (3,000) were used to calculate the observed data information.

The point estimates from the MH-RM algorithm are plotted against those from the EM algorithm in Figure 6.5. As can be seen, the estimates are reasonably close to each other. The standard errors based on the MH-RM algorithm that are obtained using the post-convergence approximation method tend to smaller than those based on the EM algorithm but reasonably compatible as expected (see Figure 6.6).

Table 6.1: Structural parameter estimates from PISA 2000 USA data analysis using the compositional effect model

	Latent variable model						Manifest variable model		
	MH-RM			EM			EM		
Parameter θ	$\hat{\theta}$	se($\hat{\theta}$)	t-value	$\hat{\theta}$	se($\hat{\theta}$)	t-value	$\hat{\theta}$	se($\hat{\theta}$)	t-value
γ_{10}	0.42	0.06	7.17	0.42	0.05	7.92	0.11	0.01	7.75
γ_{01}	0.16	0.11	1.43	0.18	0.11	1.68	0.07	0.02	3.60
τ_{00}	0.47	0.11	0.39	0.47	0.11	4.28	0.37	0.61(SD)	190.31(χ^2)
$var(\xi_{.j})$	0.12	0.07	2.30	0.11	0.06	1.86	N/A	N/A	N/A
BFLPE	-0.27	0.13	-2.12	-0.24	0.12	-1.98	-0.04	0.02	-1.76
Computation	1 hour 40 min			1 hour 40 min					
Time	M1=100, M2=300, M3=300			14qp,1 processor					
	burn-in=5								

Note1. Reported standard errors for MH-RM algorithm are from recursively approximated observed data information.
Note2. M1=Number of maximum iterations at initializing stage; M2=Number of maximum iterations at the constant gain stage; M3=Number of maximum iterations at the decreasing gain stage; qp=number of adaptive quadrature points.

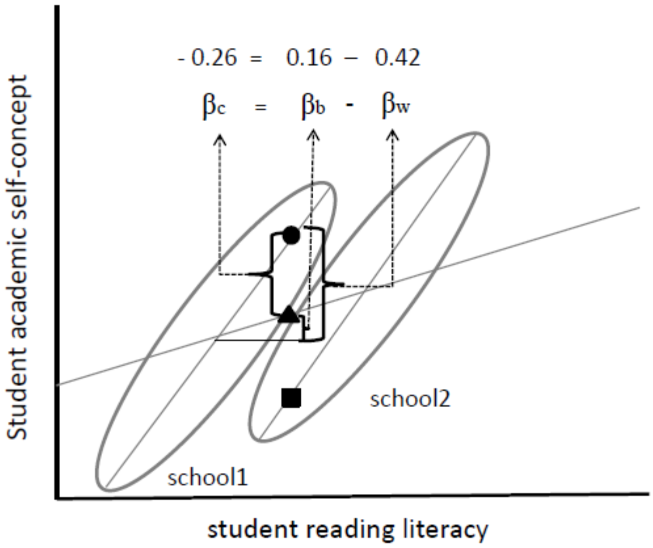


Figure 6.1: Illustration of compositional effect of academic achievement on academic self concept in literacy

Table 6.2: Item parameter estimates from PISA 2000 USA data analysis using the compositional effect model

	Slope					Threshold			
	EM		MH-RM			EM		MH-RM	
Item	$\hat{\theta}$	se($\hat{\theta}$)	$\hat{\theta}$	se($\hat{\theta}$)	Item	$\hat{\theta}$	se($\hat{\theta}$)	$\hat{\theta}$	se($\hat{\theta}$)
Y1	1.47	0.18	1.50	0.28	Y1-1	-1.15	0.15	-1.08	0.14
Y2	2.44	0.29	2.48	0.39	Y1-2	-0.09	0.13	-0.02	0.15
Y3	3.22	0.57	3.22	0.54	Y2-1	-4.58	0.40	-4.48	0.58
X1	0.83	0.10	0.83	0.11	Y2-2	-1.94	0.22	-1.83	0.30
X2	1.02	0.11	1.03	0.16	Y2-3	1.42	0.22	1.55	0.30
X3	1.40	0.17	1.40	0.25	Y3-1	-6.39	0.98	-6.20	1.14
X4	1.02	0.11	1.03	0.17	Y3-2	0.84	0.25	1.00	0.18
X5	0.89	0.10	0.90	0.11	Y3-3	0.86	0.25	1.01	0.17
X6	1.04	0.11	1.03	0.15	X1	-0.41	0.11	-0.32	0.12
X7	1.18	0.10	1.19	0.15	X2	0.32	0.12	0.42	0.13
X8	0.87	0.10	0.87	0.12	X3	1.61	0.16	1.74	0.17
X9	0.99	0.11	1.00	0.15	X4	-1.02	0.12	-0.92	0.14
X10	0.72	0.09	0.73	0.12	X5	-0.24	0.11	-0.15	0.12
X11	0.99	0.11	1.00	0.14	X6	-1.09	0.14	-0.99	0.14
X12	1.19	0.11	1.20	0.13	X7-1	0.20	0.14	0.32	0.14
X13	1.05	0.09	1.06	0.15	X7-2	1.08	0.14	1.20	0.15
X14	1.10	0.11	1.10	0.15	X8	-0.84	0.11	-0.75	0.12
X15	1.37	0.12	1.38	0.16	X9	0.75	0.12	0.85	0.12
X16	1.41	0.14	1.41	0.20	X10	0.24	0.10	0.31	0.12
X17	1.67	0.18	1.66	0.22	X11	-0.59	0.12	-0.49	0.13
X18	1.88	0.19	1.89	0.28	X12-1	-1.36	0.15	-1.24	0.15
X19	1.17	0.14	1.18	0.17	X12-2	1.72	0.15	1.84	0.17
X20	1.02	0.12	1.03	0.15	X13-1	-0.77	0.12	-0.67	0.14
X21	1.58	0.14	1.59	0.20	X13-2	2.44	0.15	2.54	0.18
X22	1.46	0.13	1.47	0.17	X14	-0.91	0.13	-0.80	0.14
X23	1.42	0.13	1.43	0.21	X15	-1.37	0.15	-1.23	0.17
X24	1.09	0.11	1.09	0.15	X16	-1.67	0.17	-1.53	0.18
X25	0.82	0.09	0.83	0.14	X17	-1.78	0.19	-1.60	0.20
X26	1.43	0.16	1.44	0.18	X18	-2.18	0.24	-1.99	0.26
X27	1.20	0.12	1.20	0.16	X19	-1.60	0.14	-1.48	0.17
X28	0.93	0.10	0.94	0.14	X20	-0.93	0.12	-0.83	0.14
X29	1.22	0.14	1.22	0.16	X21	0.20	0.16	0.36	0.17
X30	1.05	0.11	1.05	0.18	X22	0.53	0.15	0.67	0.17
X31	1.53	0.17	1.53	0.19	X23	1.64	0.18	1.79	0.19
X32	1.06	0.12	1.06	0.18	X24	-0.62	0.13	-0.51	0.13
					X25	0.52	0.11	0.60	0.12
					X26	2.14	0.20	2.29	0.27
					X27	-0.83	0.13	-0.70	0.14
					X28	-0.03	0.12	0.06	0.12
					X29	0.39	0.13	0.51	0.15
					X30	-0.71	0.13	-0.60	0.14
					X31	-1.90	0.19	-1.74	0.19
					X32	0.95	0.13	1.05	0.14

Note. One of the categories had zero frequency for the first self-concept item, so this item was analyzed by the graded response model with three categories.

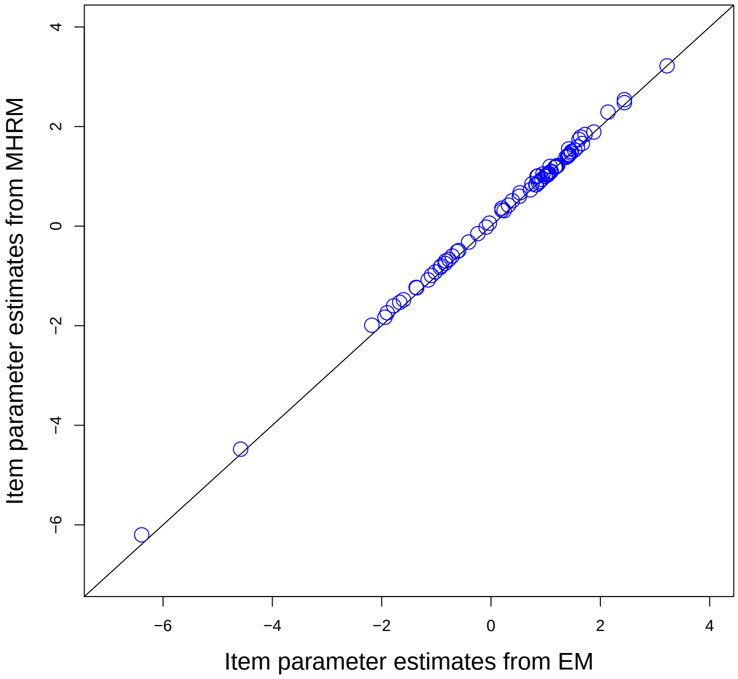


Figure 6.2: Item parameter estimates based on the EM and MH-RM algorithms, PISA 2000 USA data analysis

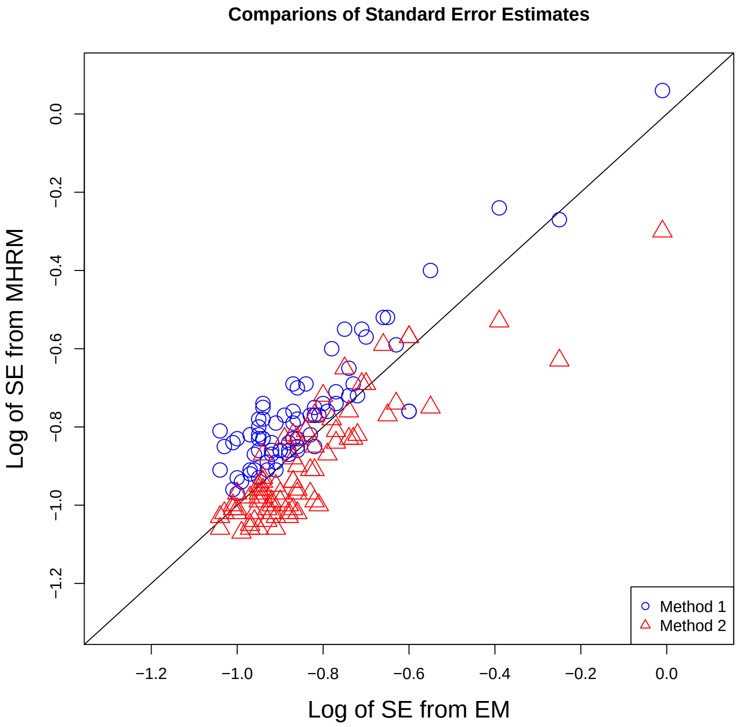


Figure 6.3: Standard errors of item parameters based on the EM and MH-RM algorithms, PISA 2000 USA data analysis. Method 1 uses recursively approximated standard errors. Method 2 uses post-convergence approximated standard errors

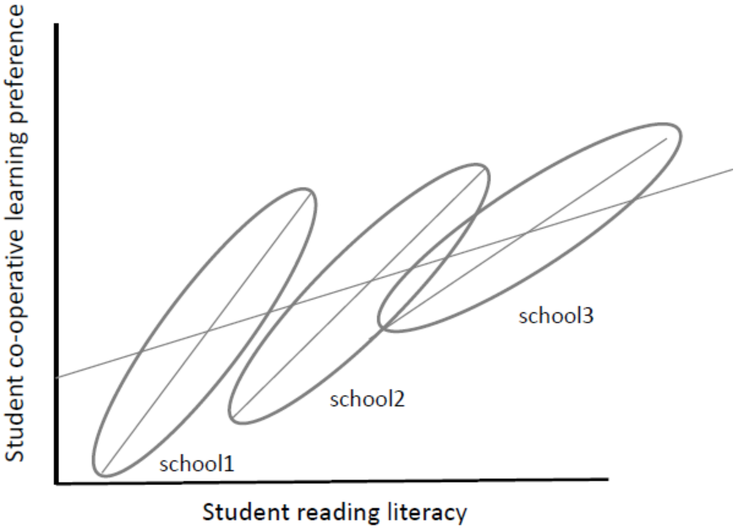


Figure 6.4: Illustration of a cross-level interaction effect of academic achievement on co-operative learning preference

Table 6.3: Structural parameter estimates from PISA 2000 Korea data analysis using the cross-level interaction model

	Latent variable model						Manifest variable model		
	MH-RM			EM			EM		
Parameter θ	$\hat{\theta}$	se($\hat{\theta}$)	t-value	$\hat{\theta}$	se($\hat{\theta}$)	t-value	$\hat{\theta}$	se($\hat{\theta}$)	t-value
γ_{10}	0.021	0.061	0.315	0.229 (0.018)	0.149	1.538	0.066	0.019	3.339
γ_{01}	0.045	0.068	0.739	0.233 (0.032)	0.009	26.972	0.041	0.016	2.618
γ_{11}	-0.088	0.062	-1.417	-0.364 (-0.050)	0.296	-1.232	-0.004	0.019	-1.363
τ_{00}	0.021	0.005	4.556	0.002 (0.034)	0.000	3.918	0.353	0.594(SD)	192.83(χ^2)
τ_{11}	0.073	0.015	4.709	1.744 (0.060)	0.615	2.837	0.005	0.070(SD)	147.04(χ^2)
τ_{01}	-0.029	0.006	-4.517	-0.052 (-0.030)	0.016	-3.211	-0.023	0.598(SD)	172.75(χ^2)
$var(\xi_{.j})$	0.817	0.007	118.852	0.629 (0.830)	0.088	7.123	N/A	N/A	N/A
Computation Time	18 hours M1=100, M2=1000, M3=1000 3000 for SE burn-in=5			8 hours 5qp,1processor Mstep iteration=5000 M convergence =0.00001					

*Note1.*Reported standard errors for the MH-RM algorithm are obtained using the post-convergence approximated observed data information.
*Note2.*Numbers in () are transformed point-estimates for comparison since different identification option was used form Mplus running.
*Note3.*M1=Number of maximum iterations at initializing stage; M2=Number of maximum iterations at the constant gain stage; M3=Number of maximum iterations at the decreasing gain stage; qp=number of adaptive quadrature points.

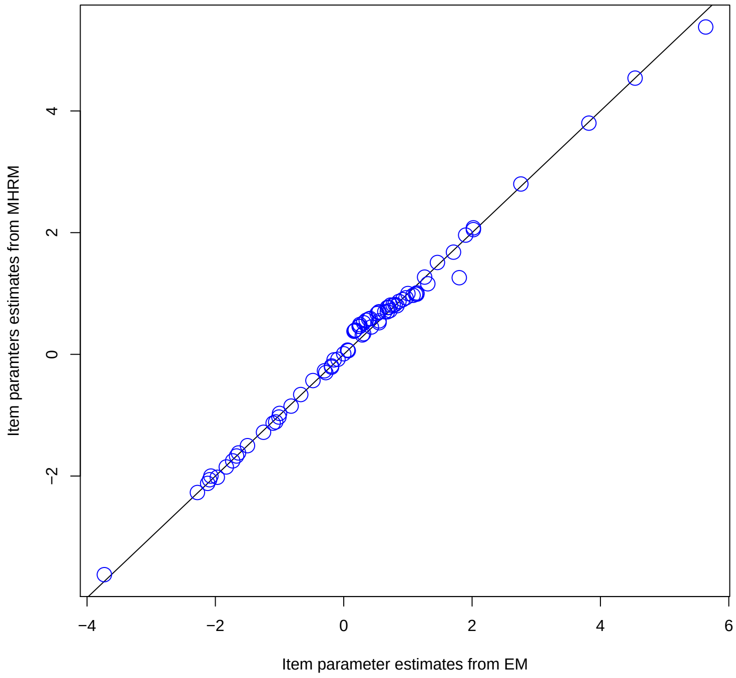


Figure 6.5: Item parameter estimates based on the EM and MH-RM algorithms, PISA 2000 Korea data analysis

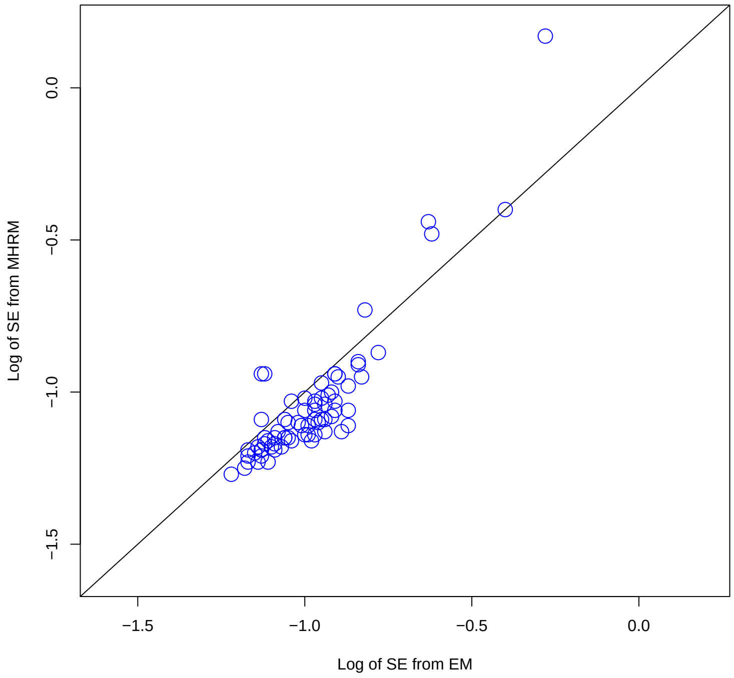


Figure 6.6: Standard errors of item parameters based on the EM and MH-RM algorithms, PISA 2000 Korea data analysis

CHAPTER 7

Discussions

7.1 Summary

Contextual effects refer to the influence of group context (group-level predictor variable) on either the level of individual actions/attitude (individual-level outcome variable) or the relationship between an individual-level outcome and an individual-level predictor. The traditional hierarchical linear modeling framework (HLM) contributed to defining these contextual effects quantitatively: the former is called a compositional effect and defined as the difference between the group-level regression coefficient and the within-group level regression coefficient, and the latter is called a cross-level interaction effect when the within-group slopes vary across groups. The particular contextual effect of interest in this study is one that occurs when a group-level characteristic of interest is measured by individual-level characteristics, and the individual-level characteristics are measured by multiple categorical indicators.

Since observed summed or averaged item scores are used for an individual level variable and observed group-means are used for a group-level variable in the traditional HLM framework, measurement error and sampling error issues have not been properly addressed. Those issues include attenuated regression coefficients and standard errors that have attracted researchers' attention. Accordingly, nonlinear multilevel latent variable modeling has been suggested as an alternative, in which latent variables are used instead of observed variables by incorporating item responses as latent variable indicators in modeling (e.g. Lüdtke et al., 2008, 2011; Marsh et al., 2009). However, a nonlinear multilevel latent variable model requires significant computation effort because the estimation process involved high dimensional numerical integration, particularly when the number of latent variables is large. This curse of dimensionality has constrained the practicability of nonlinear multilevel latent variable modeling in routine use.

The main purpose of this study was to improve estimation efficiency in obtaining full-information maximum likelihood (FIML) estimates of contextual effects by adopting the Metropolis-Hastings Robbins-Monro algorithm (MH-RM; Cai, 2008, 2010a, 2010b). R programs (R Core Team, 2012) implementing the MH-RM algorithm were produced to fit nonlinear multilevel latent variable models. Computation efficiency and parameter recovery were assessed by comparing results with an EM algorithm that uses adaptive Gauss-Hermite quadrature for numerical integration. Results indicate that the MH-RM algorithm can obtain FIML estimates and their standard errors efficiently, and the efficiency of MH-RM was more prominent for a cross-level interaction model, which requires 5-dimensional integration. While using EM algorithm with only 8 adaptive quadrature points required about 100 minutes to estimate a cross-level interaction model, the MH-RM algorithm required about 60 minutes to have similar results. Considering the difference between an interpreted language and a compiled language in which each algorithm is implemented, even more substantial improvement in efficiency is expected if the MH-RM algorithm is written in a compiled language in the future.

The second purpose of this study was to provide information about the performance of nonlinear multilevel latent variable modeling compared to traditional HLM through a simulation study with various sampling and measurement structure conditions. Results suggest that nonlinear multilevel latent variable modeling can more properly estimate and detect a contextual effect than the traditional approach in most conditions. Substantial bias was found in the between-level coefficient in the compositional model and in the cross-level interaction coefficient when the traditional model is applied, Notably, when the intraclass correlation (ICC) and the number of individuals per group were both small, the bias can be more than 80%, and the CIs hardly capture the true values. This is because that when the ICC is small, the between-group variance is too small to be decomposed and estimated, indicating between-group variation is small and the characteristic of interest is homogenous across groups. When this issue is combined with a small number of groups or a small number of people per group, the condition exacerbates the difficulty in estimating between-group variance and yield difficulty in convergence and biased estimates.

Since the within-level coefficient is also underestimated in the traditional model analysis, the point estimate of a compositional effect can be unbiased when the ICC size and the number of level-1 units per level-2 unit are both large (e.g., ICC=0.3 and the number of level-1 units per level-2 =20). However, Type I error rates of the traditional model are substantially elevated (up to 60%) in this sampling condition, indicating that the compositional effect detected by the traditional model under desirable sampling conditions could be spurious. These unacceptable Type I error rates are caused by the small standard error of between-level regression coefficient in the traditional HLM. The standard error of the between-level coefficients in HLM is influenced by the variance of between-level coefficient estimate, which is the sum of parameter dispersion and error dispersion (Raudenbush & Bryk, 2002). As the error dispersion does not reflect measurement error in HLM, the variance of between-level coefficient estimate is underestimated and so is the standard error. In contrast, the latent variable approach yielded less biased estimates, and statistical inferences across sampling and the ICC size conditions were more consistent than those of the traditional model, as long as the number of groups is sufficiently large (25 was found to be too small).

The third purpose of this study was to provide empirical illustrations using two subsets of data extracted from Programme for International Student Assessment (PISA; Adams & Wu, 2002). A negative compositional effect was found from the U.S. data in terms of the relationship between reading literacy and self-concept about reading, supporting the results from previous studies, which is called "Big-fish-little-pond" effect (e.g. Marsh et al., 2009). The compositional effect was statistically significant at $p < .05$ level when the nonlinear multilevel latent variable model was applied. On the other hand, the traditional HLM approach could not detect a statistically significant effect. It is because that the power of HLM substantially decreases when the numbers of people per group are small and this subset of data was the case. With respect to a cross-level interaction model, the relation between reading literacy and co-operative learning preference was examined, using a subset of PISA data collected in Korea. A negative, but not statistically significant, cross-level interaction was found between reading literacy and co-operative learning preference. The nonlinear multilevel latent variable model and the traditional HLM approach yielded similar results in that the cross-level interaction estimates were not statistically different from zero in both results.

Unlike the results from the simulation study, the results of empirical applications were not dramatically different in model comparison-wise. One possible explanation is that predictor variable reading literacy is measured by a large number of well-developed items for these empirical applications, and accordingly, the summed scores are very reliable. However, in other circumstances where less reliable measures (e.g., affective domain measures or teacher instructional variables) are used as predictors or where even a smaller number of people per group are sampled, it is expected to observe more substantial differences between the results from a nonlinear multilevel latent variable model and a traditional HLM. In addition, these two models also can yield divergent statistical inferences even when there are a sufficient size of ICC and a large number of people per group due the substantial elevation of Type I error rates when the traditional HLM is applied. Therefore, a wide range of further empirical applications should be followed, and the improved estimation efficiency, by adopting an MH-RM algorithm for the nonlinear multilevel latent variable models, can contribute to further applications by making the nonlinear multilevel latent variable modeling framework more practical in routine use.

7.2 Directions for Future Study

This study suggests a number of areas for further research. Above all, there is a need to make the nonlinear multilevel latent variable model more widely applicable in a wide range of research settings.

First, additional efforts are required to increase convergence rates of the cross-level interaction model. This could include exploration of the number of iterations that are needed and review of more options to approximate standard errors in an efficient and stable manner. For example, using multiple level-2 samples conditioned on a level-1 sample can be an option, but further investigation is needed to determine a proper number of samples.

Second, to give more information to users in terms of model selection, calculation of fit indices or the likelihood needs to be further investigated so that, for example, a likelihood ratio test for these two nested models can be available in the future. Having the likelihood can be especially useful in evaluating convergence of the current algorithm, since monitoring only the differences among point estimates is complicated by the scale of parameters and the size of gain constants.

Third, exploring further estimation method option is also worthwhile particularly to improve the bias in estimates when a small number of sample is used. For example, when a sample size is limited and there are too many item parameters to be estimated, two-stage estimation can be tried by using known item parameter estimates (from a scaling sample or the sample sample in typical IRT applications). There have been concerns about underestimation of standard error of latent trait in those applications, but Yang, Hansen, and Cai (2012) reported that the magnitude of underestimation is negligible when a large number of scoring sample is used and proposed a method to characterize the uncertainty in item parameter estimates. Therefore, it is expected to observe improvement in estimation efficiency as well as precision by considering a two-stage estimation approach for circumstances where scoring sample is not sufficient to estimate high dimensional model.

Fourth, an expansion of the model to multi-dimensional measurement structures such as bi-factor type model, or to structural models with more than 2-levels deserves the further research. In real data applications, items are often clustered with local dependence, and cross-level interactions can be found not only in two-level context but also in situations where 3 or more of nesting (e.g., student-teacher-school or student-school-country).

Fifth, centering of latent variables is another issue that should be addressed by multilevel latent variable model users. The latent variable model presented here imposes a group-mean centering for the level-1 variables and grand-mean centering for level-2 variables. A recent investigation reported that an un-centered or grand-mean-centered level-1 predictor produced negative bias for level-1 interaction effect, and group-mean centering produced negative bias for the level-2 interaction effect (Ryu, 2012). Accordingly, group-mean centering at level-1 and grand-mean centering at level-2 in this study seems to have been appropriate for the cross-level interaction model. Without centering, multicolinearity becomes a serious concern with the model.

Sixth, with respect to the cross-level interaction model, it will be useful to examine different options for model identification condition, providing guidance concerning whether to use a standardized factor or to anchor the first factor loadings. Theoretically, the options should yield the same results and statistical inferences. However, some differences were found when these options were tested using Mplus. This could be simply a software issue, but further exploration using different estimation approach such as MCMC with Gibbs sampler might lead to more clues about this phenomenon.

Seventh, a multiple group analysis for the compositional effect or a cross-level interaction in the framework of multilevel latent variable modeling also can be considered. In the process of empirical studies, the compositional effect appears to be different across countries (e.g., no significant compositional effect was found in the subset of Korea data). In general, further generalized multilevel latent variable modeling is required to make models more flexible to answer a broad range of questions.

Finally, The cross-level interaction model could be applied to longitudinal data (e.g., Seltzer, Choi, & Thum, 2003; Choi, Seltzer, Herman, & Yamashiro, 2007). Unlike cross-sectional data, longitudinal data are often associated with a small number of subjects (level-2 units) and a large ICC. The combination of these conditions may influence the performance of the multilevel latent variable model. Therefore, it would be worthwhile to examine the utility of this approach to various longitudinal applications.

APPENDIX A

Time-series plots of MH sampler random drawings

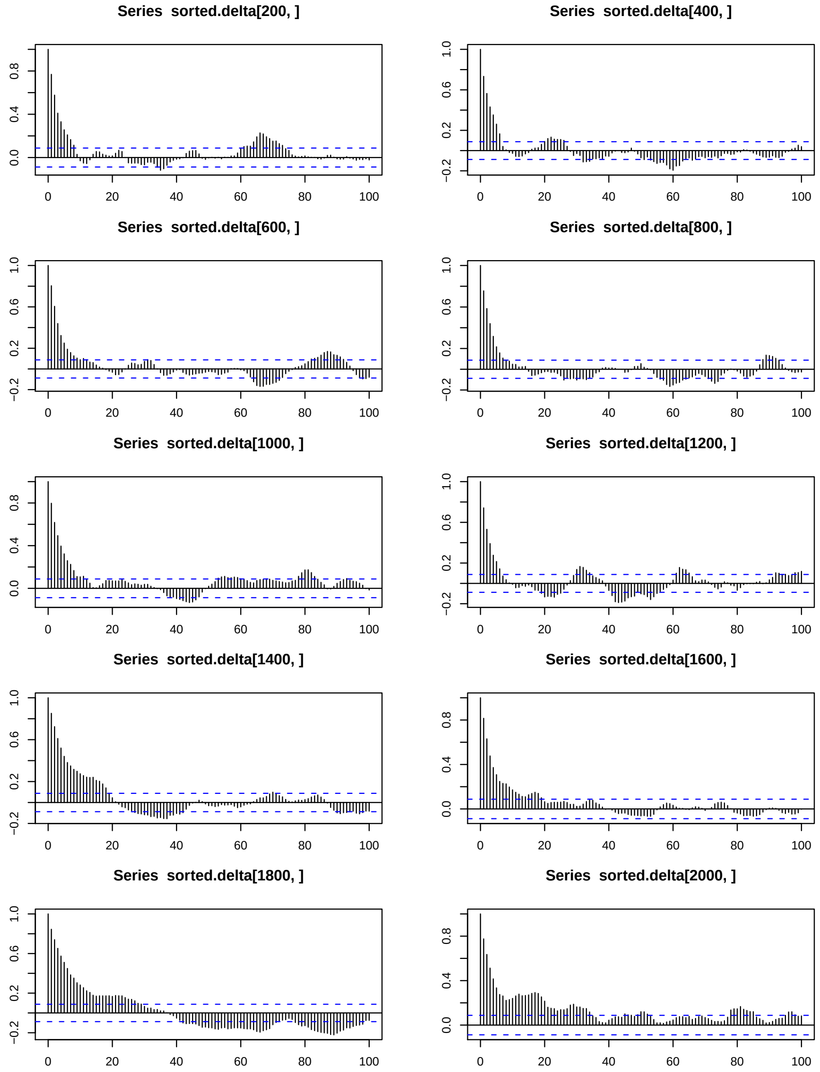


Figure A.1: The time-series plots of every 200*th* individual δ_{ij} drawings

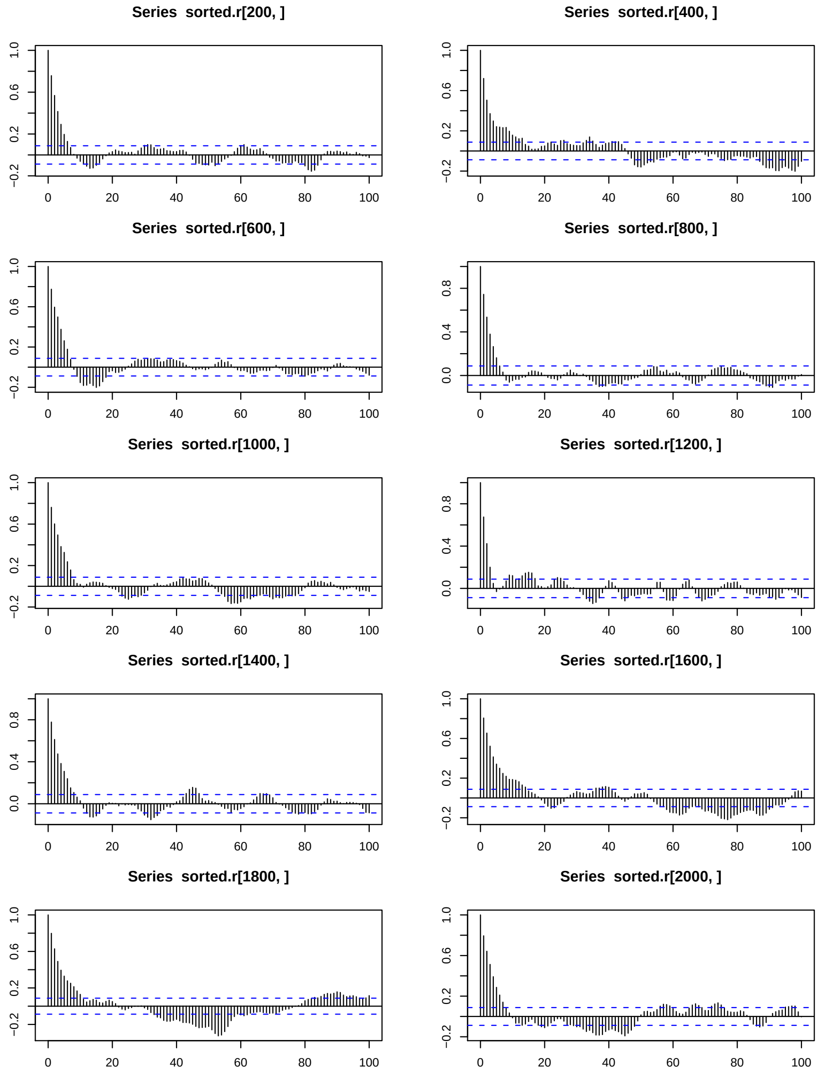
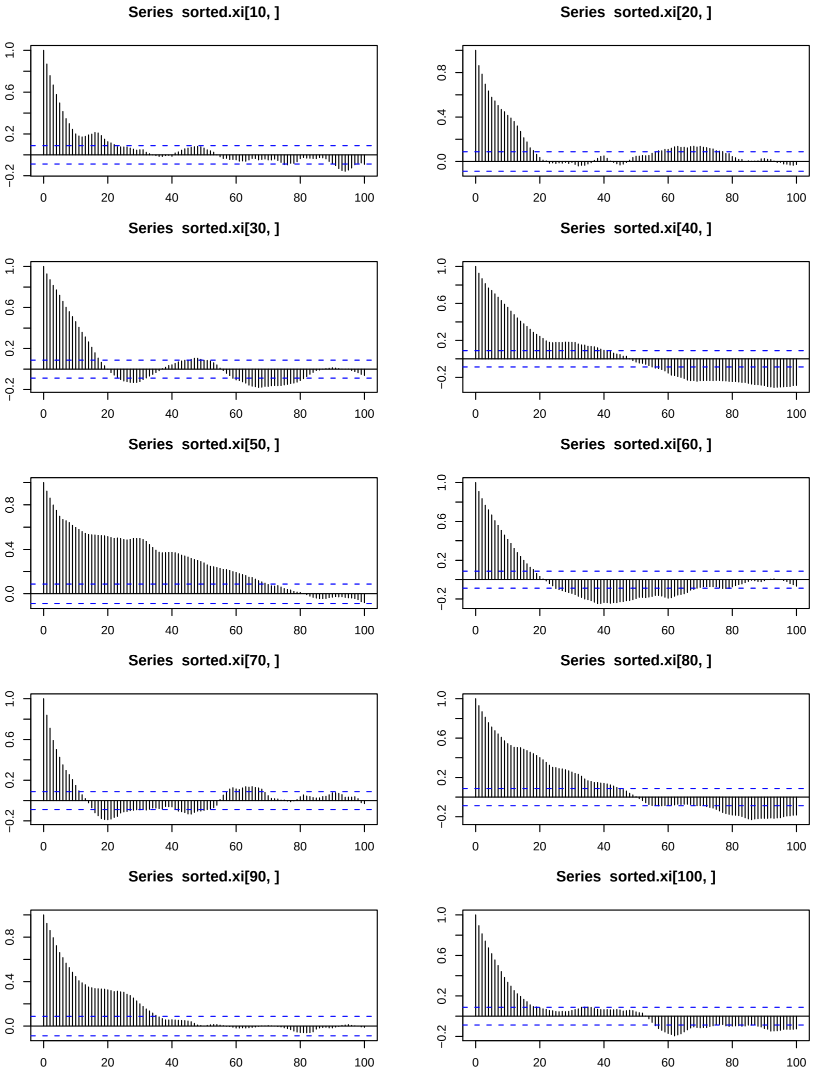


Figure A.2: The time-series plots of every 200*th* individual r_{ij} drawings


Series sorted.xi[10,]

Series sorted.xi[20,]

Series sorted.xi[30,]

Series sorted.xi[40,]

Series sorted.xi[50,]

Series sorted.xi[60,]

Series sorted.xi[70,]

Series sorted.xi[80,]

Series sorted.xi[90,]

Series sorted.xi[100,]

Figure A.3: The time-series plots of every $10th$ group $\xi_{.j}$ drawings

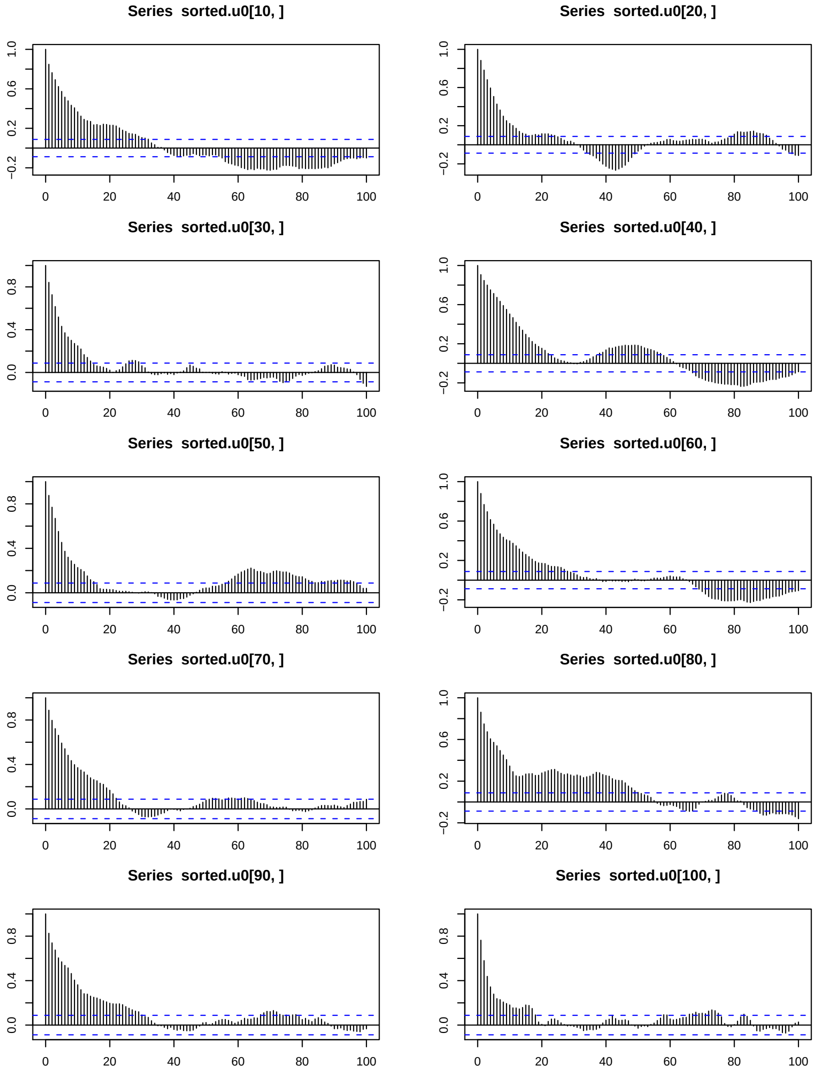


Figure A.4: The time-series plots of every 10*th* group $u0_{.j}$ drawings

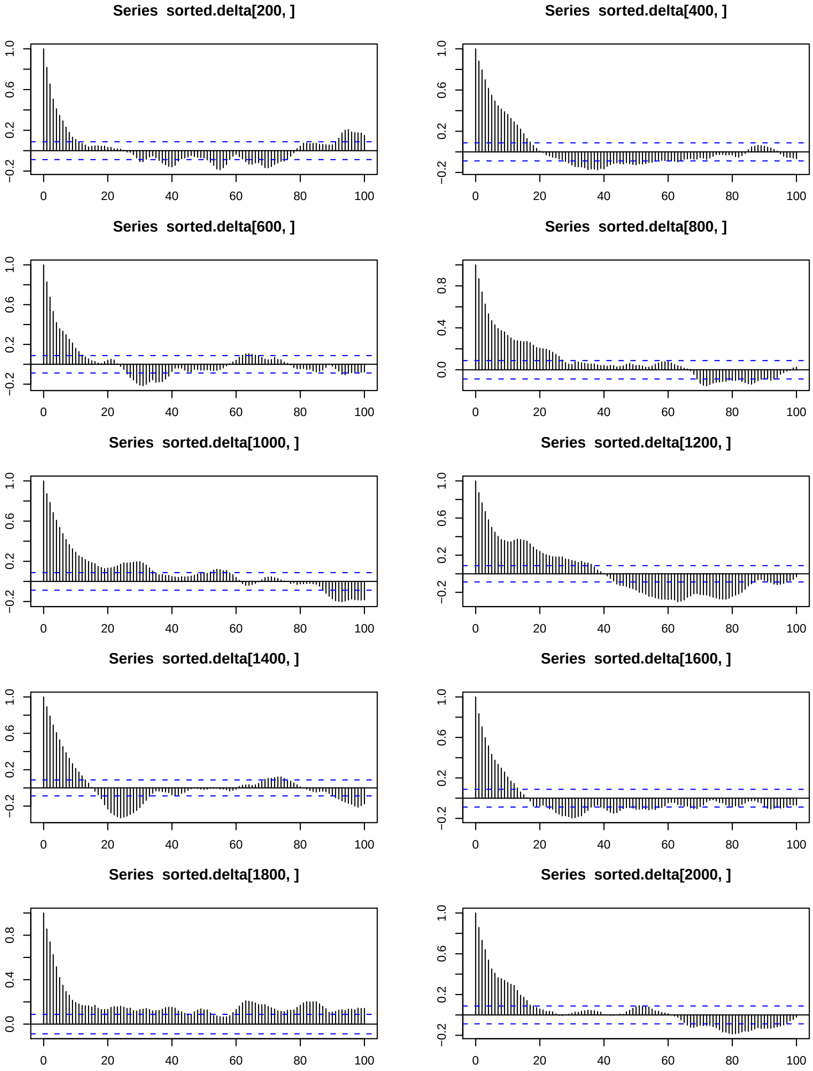


Figure A.5: The time-series plots of every 200*th* individual δ_{ij} drawings

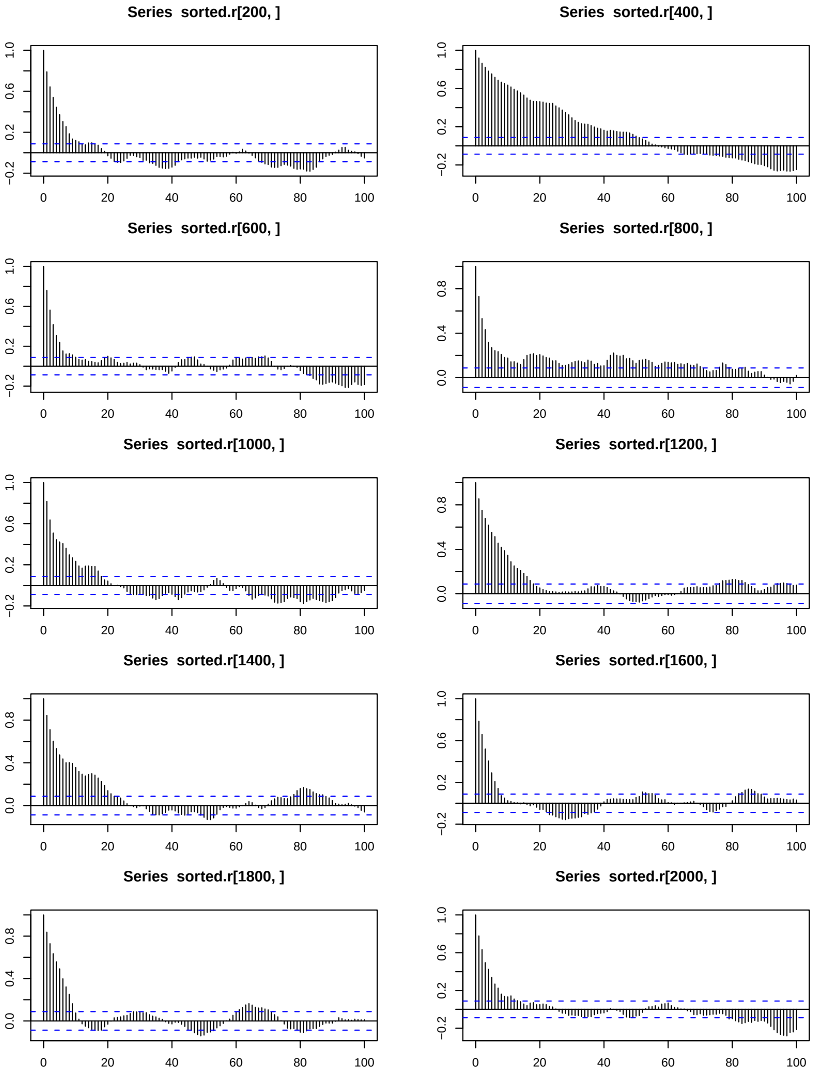


Figure A.6: The time-series plots of every 200*th* individual r_{ij} drawings

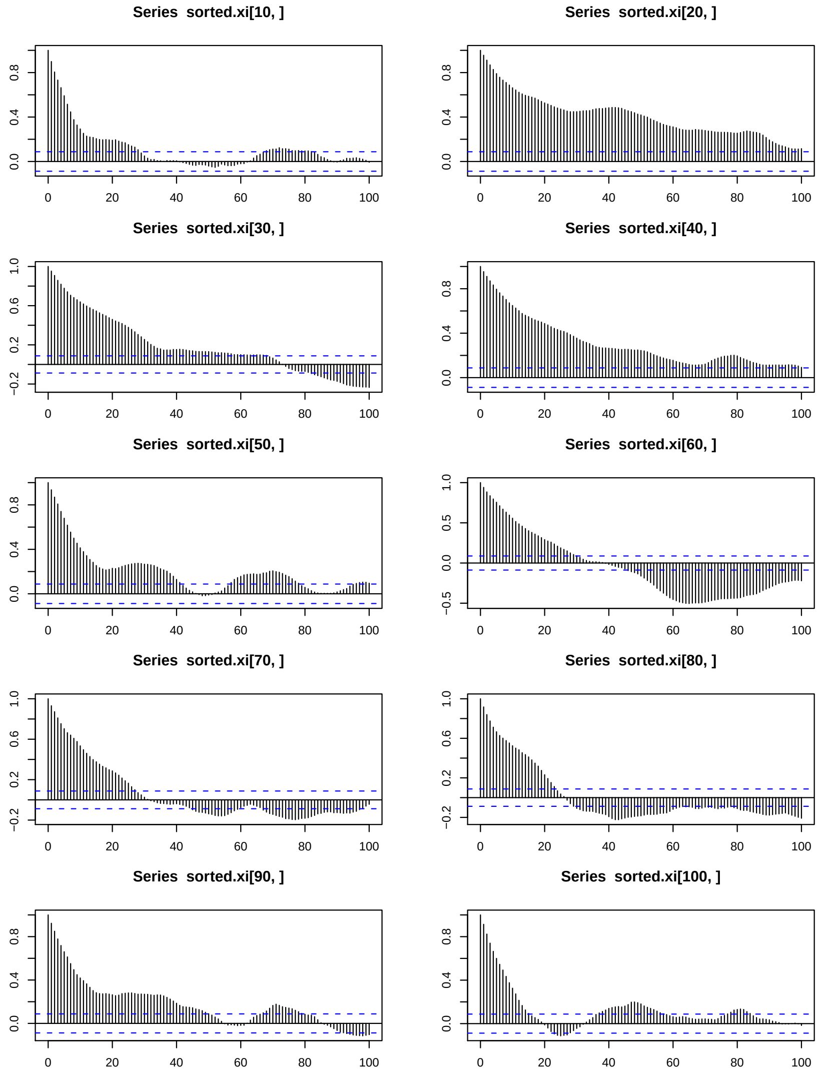


Figure A.7: The time-series plots of every 10*th* group $\xi_{.j}$ drawings

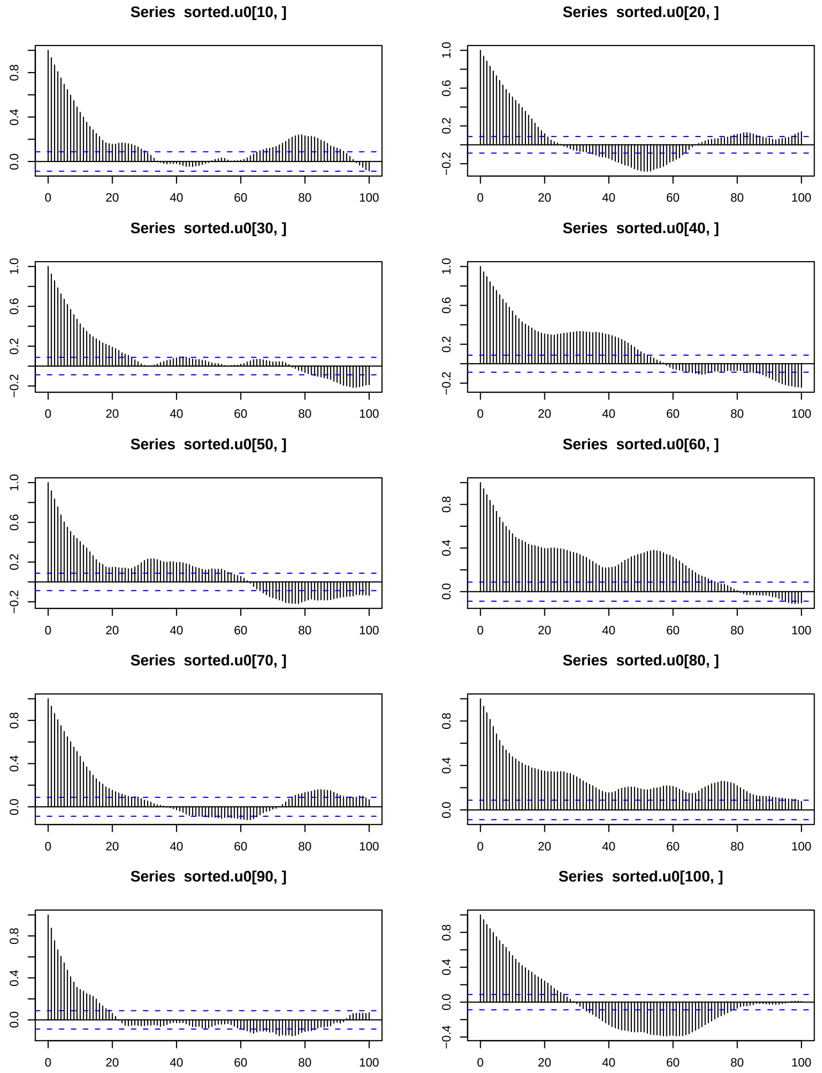
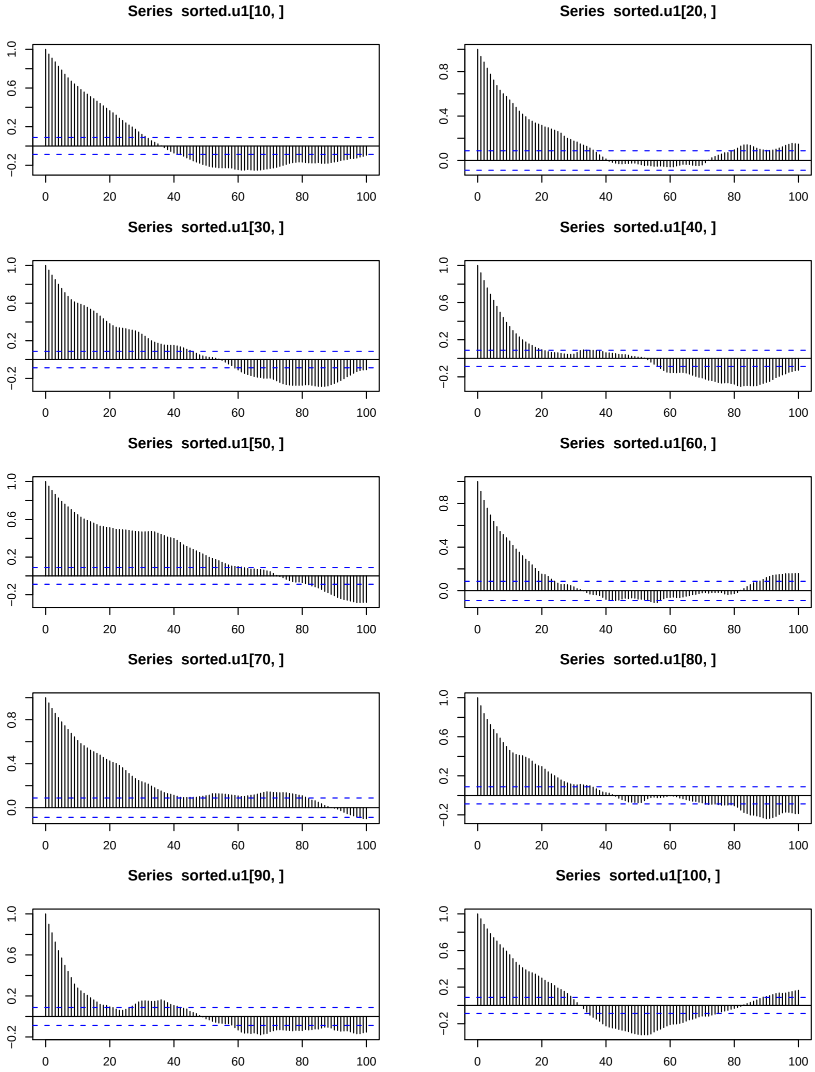


Figure A.8: The time-series plots of every 10*th* group $u0_{\cdot j}$ drawings



Series sorted.u1[10,]

Series sorted.u1[20,]

Series sorted.u1[30,]

Series sorted.u1[40,]

Series sorted.u1[50,]

Series sorted.u1[60,]

Series sorted.u1[70,]

Series sorted.u1[80,]

Series sorted.u1[90,]

Series sorted.u1[100,]

Figure A.9: The time-series plots of every $10th$ group $u1_{.j}$ drawings

Bibliography

Adams, R., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variable regression. *Journal of Educational and Behavioral Statistics*, *22*, 47-76.

Adams, R., & Wu, M. (2002). *PISA 2000 technical report*. Paris, France: Organization for Economic Cooperation and Development.

Allen, M. J., & Yen, W. M. (2001). *Introduction to Measurement Theory*. Long Grove, IL: Waveland Pr, Inc.

Ansari, A., & Jedidi, K. (2000). Bayesian factor analysis for multilevel binary observations. *Psychometrika*, *65*, 475-496.

Arminger, G., & Muthén, B. (1998). A Bayesian approach to nonlinear latent variable models using the Gibbs sampler and the Metropolis-Hastings algorithm. *Psychometrika*, *63*, 271-300.

Baker, T., & Clark, J. (2010). Cooperative learning a double edged sword: A cooperative learning model for use with diverse student groups. *Intercultural Education*, *21*, 257-268.

Bauer, D. J., & Cai, L. (2009). Consequences of unmodeled nonlinear effects in multilevel models. *Journal of Educational and Behavioral Statistics*, *34*, 97–114.

Bentler, P. M. (1985). Theory and implementation of EQS: A structural equations program. *Sociological Methods & Research*, *16*(1), 78-117.

Bliese, P. D. (2000). Multilevel theory, research, and methods in organizations. In K. J. Klein & S. W. Kozlowski (Eds.), (p. 349-381). San Francisco: Jossey-Bass.

Bliese, P. D., Chan, D., & Ployhart, R. E. (2007). Multilevel methods: Future directions in measurement, longitudinal analyses, and nonnormal outcomes. *Organizational Research Methods*, *10*, 551-563.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: John Wiley & Sons.

Bottoms, A. E., & Wiles, P. (2004). *The Oxford Handbook of Criminology* (R. Morgan &

R. Reiner, Eds.). Oxford: Clarendon Press.

Brennan, R. L. (1992). Generalizability Theory. *Educational Measurement: Issues and Practice*, *11*(4), 27-34.

Cai, L. (2008). *A Metropolis-Hastings Robbins-Monro Algorithm for Maximum Likelihood Nonlinear Latent Structure Analysis with a Comprehensive Measurement Model*. Unpublished doctoral dissertation, Department of Psychology, University of North Carolina - Chapel Hill.

Cai, L. (2010a). High-dimensional exploratory item factor analysis by a Metropolis-Hastings Robbins-Monro algorithm. *Psychometrika*, *75*, 33–57.

Cai, L. (2010b). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, *35*, 307–335.

Celeux, G., Chauveau, D., & Diebolt, J. (1995). *On stochastic versions of the EM algorithm* (Tech. Rep. No. 2514). The French National Institute for Research in Computer Science and Control.

Celeux, G., & Diebolt, J. (1991). *A stochastic approximation type EM algorithm for the mixture problem* (Tech. Rep. No. 1383). The French National Institute for Research in Computer Science and Control.

Choi, K., Seltzer, M., Herman, J., & Yamashiro, K. (2007). Children Left Behind in AYP and Non-AYP schools: Using Student Progress and the Distribution of Student Gains to Validate AYP. *Educational Measurement: Issues and Practice*, *26*(3), 21-32.

Coleman, J., Hoffer, T., & Kilgore, S. (1982). Cognitive outcomes of public and private schools. *Sociology of Education*, *58(2/3)*, 65-76.

Cronbach, L., Gleser, G. C., Nanda, H., & Rajaratham, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for socres and profiles*. New York: John Wilery and Sons.

Croon, M. A., & van Veldhoven, M. J. (2007). Predicting group level variables from variables measured at the individual level: A latent variable multilevel model. *Psychological Methods*, *12*, 45-57.

Cudeck, R., & du Toit, S. H. C. (2003). Multilevel modeling: Methodological advances, issues and applications. In N. Duan & S. P. Reise (Eds.), (p. 1-24). Mahwah, NJ: Erlbaum.

Cudeck, R., Harring, J. R., & du Toit, S. H. C. (2009). Marginal maximum likelihood estimation of a latent variable model with interaction. *Journal of Educational and Behavioral Statistics*, *34*, 131–144.

de Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer.

Delyon, B., Lavielle, M., & Moulines, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *The Annals of Statistics*, *27*, 94-128.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society*, *39*, 1-38.

Dempster, A. P., Rubin, D. B., & Tsutakawa, R. K. (1981). Estimation in covariance components models. *Journal of the American Statistical Association*, *76*, 341–353.

Elston, R. C., & Grizzle, J. E. (1962). Estimation of time response curves and their confidence bands. *Biometrics*, *18*, 148-159.

Erbring, L., & Young, A. (1979). Individuals and social structure: Contextual effects as endogenous feedback. *Sociological Moethods and Research*, *7*, 396-430.

Firebaugh, G. (1978). A rule for inferring individual level relationships from aggregate data. *American Sociological Review*, *43*, 557-572.

Fisher, R. A. (1925). Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society*, *22*, 700-725.

Fox, J. P., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika*, *66*, 269-286.

Fuller, W. A. (1987). *Measurement error models*. New York: John Wiley.

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*, 398-409.

Gelman, A., Gilks, W. R., & Roberts, G. O. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability*, *7*(1), 110-120.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 721-741.

Goldstein, H., Bonnet, G., & Rocher, T. (2007). Multilevel structural equation models for the analysis of comparative data on educational performance. *Journal of Educational and Behavioral Statistics*, *32*(3), 252-286.

Goldstein, H., & Browne, W. (2004). Multilevel factor analysis models for continuous and discrete data. In Maydeu-Olivares & M. J. J. (Eds.), *Contemporary psychometrics* (p. 7270-7274). Mahwah, NJ: Lawrence Erlbarum Associates.

Hastings, W. K. (1970). Monte carlo simulation methods using markov chains and their applications. *Biometrika*, *57*, 97-109.

Henry, K. L., & Slater, M. D. (2007). The Contextual Effect of School Attachment on Young Adolescents' Alcohol Use. *Journal of School Health*, *77*(2), 67–74.

Iversen, G. R. (1991). *Contextual analysis*. Newbury Park, CA: Sage.

Johnson, D. W., & Johnson, R. T. (1989). *Cooperation and competition: Theory and research*. Edina, MN: Interaction Book Company.

Jöreskog, K. G. (1973). Structrual equation models in the social sciences. In A. S. Goldberger & O. D. Duncan (Eds.), (p. 85-112). New York: Academic Press.

Jöreskog, K. G., & Sörbom, D. (1974). *LISREL III [computer software]*. Chicago, IL. (Computer Software)

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement*, *38*, 79-93.

Kamata, A., Bauer, D. J., & Miyazaki, Y. (2008). Multilevel modeling of educational data. In A. A. OConnell & M. D. B. (Eds.), (pp. 345–388). Charlotte, NC: Information Age Publishing.

Keesling, J. W. (1972). *Maximum likelihood approaches to causal analysis*. Unpublished doctoral dissertation, Department of Education, University of Chicago.

Kenny, D. A., & Judd, C. M. (1984). Estimating the nonlinear and interactive effects of latent variables. *Psychological Bulletin*, *96*, 201-210.

Kozlowski, S. W. J., & Klein, K. J. (2000). Multilevel theory, research, and methods in organizations. In K. J. Klein & S. W. J. Kozlowski (Eds.), (p. 3-90). San Francisco: Jossey-Bass.

LaHuis, D. M., & Ferguson, M. W. (2009). The accuracy of significance tests for slope variance components in multilevel random coefficient models. *Organizational Research Methods*, *12*, 418-443.

Laird, N. M., & Ware, H. (1982). Random-effects models for longitudinal data. *Biometrics*, *38*, 963-974.

Lee, S. Y. (1990). Multilevel analysis of structural equation models. *Biometrika*, *77*, 763-772.

Lee, S. Y., & Poon, W. Y. (1998). Analysis of two-level structural equation models via EM type algorithms. *Statistica Sinica*, *8*, 749-766.

Lee, S. Y., Song, X. Y., & Poon, W. Y. (2004). Comparison of approaches in estimating interaction and quadratic effects of latent variables. *Multivariate Behavioral Research*, *39*, 37-67.

Lee, V. E., & Bryk, A. (1989). A multilevel model of the social distribution of educational achievement. *Sociology of Education*, *62*, 172-192.

Liang, J., & Bentler, P. M. (2004). An EM algorithm for fitting two-level structural equation models. *Psychometrika*, *69*, 101-122.

Longford, N. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced modles with nested random effects. *Biometrika*, *74*(4), 817–827.

Longford, N. (1993). *Random coefficients models*. Oxford: Clarendon.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Louis, T. A. (1982). Fiding the Observed Information Matrix when Using the EM Algorithm. *Journal of the Royal Statistical Society*, *44*(2), 226-233.

Lüdtke, O., Marsh, H., Robitzsch, A., & Trautwein, U. (2011). A 2 x 2 taxonomy of multilevel latent covariate models: Accuracy and bias trade-offs in full and partial error-correction models. *Psychological Methods*, *16*(4), 444-467.

Lüdtke, O., Marsh, H., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods*, *13*, 203-229.

Maier, K. (2001). A Rasch hierarchical measurement model. *Journal of Educational and Behavioral Statistics*, *26*, 307-330.

Marsh, H. W. (1987). The big-fish-little-pond effect on academic self concept. *Journal of Educational Psychology*, *79*, 280-295.

Marsh, H. W., Lüdtke, O., Robitzsch, A., Trautwein, U., Asparouhov, T., Muthén, B., & Nagengast, B. e. a. (2009). Doubly-latent models of school contextual effects: Integrating multilevel and structural equation approaches to control measurement and sampling error. *Multivariate Behavioral Research*, *44*, 764-802.

Mason, W. M., Wong, G. Y., & Entwisle, B. (1983). *Sociological methodology* (S. Leinhardt, Ed.). San Francisco: Jossey-Bass.

McDonald, R. P. (1993). A general model for two-level data with responses missing at random. *Psychometrika*, *58*, 575-585.

McDonald, R. P. (1994). The Bilevel Reticular Action Model for path analysis with latent variables. *Sociological Methods & Research*, *22*, 399-413.

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*, 215-232.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equations of state space calculations by fast computing machines. *Journal of Chemical Physics*, *21*, 1087-1092.

Muthén, B. O. (1990). Mean and covariance structure analysis of hierarchical data. *Journal of Educational Measurement*, *28*, 338-354.

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement*, *28*, 338-354.

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods &*

Research, *22*, 376-398.

Muthén, L. K., & Muthén, B. O. (2008). Mplus 5.0 [Computer software]. Los Angeles, CA.

Muthén, L. K., & Muthén, B. O. (2010). *Mplus user's guide* (Sixth ed.). Los Angeles, CA: Muthèn & Muthèn.

Oberwittler, D. (2004). A multilevel analysis of neighbourhood contextual effects on serious juvenile offending: The role of subcultural values and social disorganization. *European Journal of Criminology*, *1*(2), 201-235.

Owens, L., & Barnes, J. (1992). *Learning Preferences Scales* (Tech. Rep.). Hawthorn, Vic.: Australian Council for Educational Research.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team. (2012). nlme: Linear and nonlinear mixed effects models [Computer software manual]. (R package version 3.1-104)

R Core Team. (2012). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org` (ISBN 3-900051-07-0)

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika*, *69*, 167-190.

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

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Raudenbush, S. W., & Willms, J. (1995). The estimation of school effects. *Journal of Educational and Behavioral Statistics*, *20(4)*, 307-335.

Robbins, H., & Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, *22*, 400-407.

Rosenberg, B. (1973). Linear regression with randomly dispersed parameters. *Biometrika*, *60*, 61-75.

Ryu, E. (2012). *Interaction of level-1 variables in multilevel structural equation models.* Unpublished paper presented atInternational Meeting of the Psychometric Society, Lincoln, Nebraska.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monographs, 17*.

Sampson, R. J., Morenoff, J. D., & Gannon-Rowley, T. (2002). Assessing neighborhood effects: Social processes and new directions in research. *Annual Review of Sociology, 28*, 443-78.

Seltzer, M., Choi, K., & Thum, Y. M. (2003). Examining relationships between where students start and how rapidly they progress: Using new developments in growth modeling to gain insight into the distribution of achievement within schools. *Educational Evaluation and Policy Analysis, 25*, 263-286.

Singer, J. D. (1998). Using SAS PROC MIXED to multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statstics, 23*(4), 323-355.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models.* Boca Raton, FL: Chapman and Hall/CRC.

Spearman, C. (1904). "General intelligence" objectively determined and measured. *American Journal of Psychology, 15*, 201-293.

Tsay, M., & Brady, M. (2010). A case study of cooperative learning and communication pedagogy; does working in terms make a difference? *Journal of the Scholarship of Teaching and Learning, 10*, 78-89.

Wikström, P.-O. H. (1998). *Oxford handbook on crime and punishment* (M. Tonry, Ed.). Oxford: Oxford University Press.

Wiley, D. E. (1973). Structrual equation models in the social sciences. In A. Goldberger & O. D. Duncan (Eds.), (p. 69-83). New York: Academic Press.

Willms, J. D. (1986). Social class segregation and its relationship to pupils' examination results in scotland. *American Socialological Review, 55*, 224-241.

Wooldredge, J. D., & Thistlethwaite, A. (1999). *Reconsidering Domestic Violence Recidivism:*

Individual and Contextual Effects of Court Dispositions and Stake in Conformity (Tech. Rep.). Cincinnati, OH: University of Cincinnati. (Final Report submitted to the National Institute of Justice)

Wright, S. (1918). On the nature of size factors. *Genetics*, *3*, 367-374.

Wright, S. (1921). Correlation and Causation. *Journal of Agricultural Research*, *20*, 557-585.

Wright, S. (1934). The method of path coefficients. *Annals of Mathmatical Statistics*, *5*, 161-215.

Wright, S. (1960). Path Coefficients and path regressions: Alternative or complementary concepts? *Biometrics*, *16*, 189-202.

Yang, J. S., Hansen, M., & Cai, L. (2012). Characterizing sources of uncertainty in item response theory scale scores. *Educational and Psychological Measurement*, *72*(2), 264-290.