Modeling global and local person dependence for clustered samples in Rasch models

by

JIN, Kuan-Yu

A Thesis Submitted to

The Education University of Hong Kong

in Partial Fulfillment of the Requirement for

the Degree of Doctor of Philosophy

June 2017

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

Statement of Originality

I, JIN, Kuan-Yu, hereby declare that I am the sole author of the thesis and the material presented in this thesis is my original work except those indicated in the acknowledgement. I further declare that I have followed the University's policies and regulations on Academic Honesty, Copyright and Plagiarism in writing the thesis and no material in this thesis has been submitted for a degree in this or other universities.

_____

JIN, Kuan-Yu

June 2017

Thesis Examination Panel Approval

Members of the Thesis Examination Panel approved the thesis of JIN, Kuan-Yu defended on 15 May 2017.

Principal Supervisor
Prof WANG, Wen-Chung
Chair Professor
Department of Psychological Studies
The Education University of Hong Kong

External Examiner
Prof CHIU, Ming Ming
Professor
College of Education
Purdue University, USA

Associate Supervisor
Prof MOK, Magdalena Mo Ching
Chair Professor
Department of Psychological Studies
The Education University of Hong Kong

External Examiner
Dr. YU, Hsiu-Ting
Associate Professor
Department of Psychology
National Chengchi University, Taiwan

Associate Supervisor
Dr KWAN, Lok Yin Joyce
Assistant Professor
Department of Psychological Studies
The Education University of Hong Kong

External Examiner
Prof LEI, Pui-Wa
Professor
Educational Psychology, Counseling, and Special Education
Penn State University, USA

Approved on behalf on the Thesis Examination Panel:

_____

Chair, Thesis Examination Panel
Prof Lo, Sing Kai
Chair Professor
Dean of Graduate School
The Education University of Hong Kong

Abstract

Cluster sampling is widely applied in social science. Respondents recruited from the same clusters may behave more similarly than those from different clusters in terms of their general proficiency, as well as their response patterns. The homogeneity of general proficiency refers to global person dependence (GPD), which can be adequately accounted for by means of multilevel modeling. Local person dependence (LPD) describes some kinds of interpersonal interactions that are conditional on respondents' proficiency levels, implying that person residuals are not locally independent when fitting a standard (e.g., Rasch) model. Many item response theory (IRT) models have been developed to create a multilevel structure for describing GPD, but few address the occurrence and influence of LPD.

This study was intended to develop a new class of Rasch models for clustered samples to account for GPD and LPD jointly so that the two kinds of dependence can be quantified. In brief, I developed a new set of IRT models for integrating multilevel structures on the measured latent trait(s) and a component of random item difficulty across person clusters. The simple models for dichotomous and polytomous responses were displayed in sequence, and the extensions to multiple tests and many-faceted data were illustrated on top of the basic forms. These models can be easily implemented by means of WinBUGS, a freeware application used for Bayesian analysis. A series of simulations were carried out to examine the parameter recovery of the new models, as well as the consequences of fitting standard models without considering LPD. The results indicated that the parameters of the new models can be recovered very well, and that ignoring LPD by fitting standard models elicits biased estimation and inflated GPD.

The technique of cluster analysis is to group subjects in accordance with the homogeneity among a set of variables. Therefore, it may be helpful to assess the occurrence of LPD, and how respondents within a cluster are grouped together, especially when the

magnitude of LPD is substantial. The effectiveness of hierarchical cluster analysis (HCA) in exploring the dependence of person residuals was examined. It was found that HCA was useful for recovering respondents' true membership by means of the homogeneity information among person residuals, but it was not always sensitive to LPD.

Four empirical examples – the National Longitudinal Study of Adolescent Health (Add Health) project, the Impact of Community Policing Training and Program Implementation on Police Personnel in Arizona study, the International Civic and Citizenship Study in 2009, and the Love Relationship Scale for couples – were used to demonstrate the new models. In the first and second examples particularly, items were designed to measure a single latent trait, whereas in the third and fourth examples, items were assembled as different subtests measuring distinct, but correlated, latent traits. It was found that, in these four examples, the clustered samples exhibited various degrees of LPD on items. As for the findings in the simulations, fitting simpler models without regarding the influence of LPD yielded shrunken scales and inflated GPD.

Finally, conclusions were drawn based on the findings, in which the importance of the consideration of LPD when dealing with clustered samples was emphasized, and the implications of how to interpret LPD were discussed. Limitations in the LPD modeling approach and in HCA for accessing LPD also were addressed. Suggestions for future studies also were provided.

Keywords: clustered sample, local person dependence, Rasch models, multidimensional item response theory.

Acknowledgement

Over the past seven years in Hong Kong, I have received support and encouragement from a great number of individuals. I would like to express my deep sense of gratitude to my principal supervisor, Prof. Wen-Chung Wang, who has been a patient mentor and a good friend. His guidance has made this an interesting and rewarding journey. I also would like to thank my dissertation committee of Professor Magdalena Mo Ching Mok and Dr. Lok Yin Joyce Kwan for their assistance as I moved from an idea to a completed study. In addition, I would like to thank Dr. Hui-Fang Chen for her valuable and helpful advice and support.

My appreciation extends to all members of the Assessment Research Centre for their academic and administrative support, especially Sheng-Yun Huang, Chia-Ling Hsu, Nicky Li, Chen-Wei Liu, Jingjing Yao, Sze Ming Lam, Jinxin Zhu, and Kun Xu. I also would like to thank my schoolmates at the Education University of Hong Kong, including Joseph Chow, Iceman Leung, Li Liu, and Lijuan Li for their encouragement throughout my research. In addition, I want to express my appreciation to my girlfriend, Yiting Liu, who stuck with me during my long months of writing and revision.

Finally, I would like to thank all my family members for their love and constant encouragement and motivation. The thesis is heartily dedicated to my mother, who took the lead to heaven.

Table of Contents

List of Abbreviations

AIC	Akaike information criterion
BIC	Bayesian information criterion
CDM	Cognitive diagnostic model
DIC	Deviance information criterion
DIF	Differential item functioning
DRF	Differential rater functioning
ESS	Effective sample size
GPD	Global person dependence
HCA	Hierarchical cluster analysis
HOPCM	Higher-order partial credit model
HOPCM-CS	Higher-order partial credit model for clustered samples
ICC	Intra-class coefficient
ICCS	International Civic and Citizenship Study
IELTS	International English Language Testing System
IRT	Item response theory
LID	Local item dependence
LPD	Local person dependence
M2PLM-CS	Multilevel two-parameter logistic model for clustered samples
MCMC	Markov chain Monte Carlo
MDMPCM	Multidimensional multilevel partial credit model
MDMPCM-CS	Multidimensional multilevel partial credit model for clustered samples
MPCM	Multilevel partial credit model
MPCM-CS	Multilevel partial credit model for clustered samples
MFRM	Many-faceted Rasch model

MFRM-CS	Many-faceted Rasch model for clustered samples
MGPCM-CS	Multilevel generalized partial credit model for clustered samples
MRM	Multilevel Rasch model
MRM-CS	Multilevel Rasch model for clustered samples
PCM	Partial credit model
PISA	Program for International Student Assessment
RDSM	Response dependence of subjects model
RMSE	Root mean square error
RSM	Rating scale model

List of Figures

List of Tables

Chapter 1: Introduction

1.1 Motivation

Item response theory (IRT) models are developed to conduct categorical responses and have been widely applied in academic disciplines, including education, psychology, sports, and marketing. Local independence is one of the important assumptions in IRT models. There are two major kinds of violations of the assumption: local item dependence (LID) and local person independence (LPD). In LID, item residuals in a test are dependent, and in LPD, person residuals in a test are also dependent. If the assumption of local independence is violated, fitting a standard model would result in biased parameter estimates and a misleading conclusion (Yen, 1993). Many factors that may cause LID have been investigated, including testlet or item-bundle structures (Wainer, Bradlow, & Wang, 2007; Wang & Wilson, 2005), negatively worded items (Wang, Chen, & Jin, 2015), and non-ignorable missingness (Glas & Pimentel, 2008; Holman & Glas, 2005), among others. Many studies (Chen & Wang, 2007; Tuerlinckx, & De Boeck, 2001; Wang & Jin, 2016) have shown that test reliability—equivalently, the precision of measurement—can be seriously inflated or deflated when the magnitude of violation is substantial. Compared with the great attention that has been paid to LID, the effect of LPD has not been properly investigated so far.

Let's say you have a two-dimensional data matrix $\mathbf{Y}$ summarizing N persons responding to I items with J categories. Also, $y_{nij} = 1$ if person n gets score j on item i; otherwise, $y_{nij} = 0$. In this case, the likelihood function, which is the simple product of the probabilities for all responses, can be formulated as the following:

$$L(\mathbf{Y} \mid \boldsymbol{\theta}, \boldsymbol{\delta}) = \prod_{n=1}^{N} \prod_{i=1}^{I} \prod_{j=1}^{J} (P_{nij})^{y_{nij}} , \quad (1)$$

in which P_{nij} is the probability of endorsing score j for person n on item I, and $\boldsymbol{\theta}$ and $\boldsymbol{\delta}$ are person and item parameters, respectively. Based on the likelihood function, the estimates of

person and item parameters can be derived by means of maximum likelihood estimation or Bayesian inference. It should be noted that the expression of the likelihood function is valid only if the local independence assumption is met, i.e., when person and item parameters are unable to fully account for the interrelationships among items (or persons), suggesting the existence of LID (or LPD), in which case, the nuisance would interfere with the estimation, and biased parameter estimates may be generated.

1.2 Global and local dependence

In this paper, the general terminology of dependence refers to two specific notions: global dependence and local dependence. Because item responses collected through tests or instruments are usually two-faceted data, including item characters and person latent traits, the concepts of item dependence and person dependence are sequentially introduced in this section.

Global item dependence, which is also realized as the internal consistency of a test, refers to the dependence of item scores between items. Because items assembled in a test are designed to measure the common construct, they are expected to be homogeneous in their contents (Wilson, 2004). Global item dependence can be quantified by computing the Cronbach's α or the inter-item correlations, and a high value is usually demanded. In most cases, practitioners are reluctant to accept a low value of Cronbach α or averaged inter-item correlations, as this implies that the items are too divergent to measure the target proficiency. In the case of low global item dependence, one should review the original items carefully and, when necessary, remove inappropriate items from the test.

Comparatively, LID refers to the idea that, after conditioning on item and person parameters, item residuals are not purely random errors and still relate to each other. The existence of LID suggests that, in addition to the intended-to-be-measured latent trait, there is some covariation among item responses. For instance, testlet-based items, which are linked

by a common stimulus (e.g., a reading passage or a figure), are widely applied in educational and psychological tests. It has been acknowledged that items within the same testlet may not be locally independent because knowing or not knowing the answer to an item may influence the chances of success on other items in the same testlet (Yen, 1993). Under such a case, fitting standard IRT models will result in biased parameter estimates (Sireci, Thissen, & Wainer, 1991; Wainer, Bradlow, & Du, 2000; Wainer, Bradlow, & Wang, 2007; Wang & Wilson, 2005). Another example is the wording effect. In the social sciences, to minimize acquiescence in responding to rating-scale items, inventories are usually a mix of positively and negatively worded items (DeVellis, 2005; Kieruj & Moors, 2013). In most cases, responses to negatively worded items are coded in reverse, so that all items on the scale are presumed to go in the same wording direction. A series of experiments, however, showed that responding to positively and negatively worded items involves two different cognitive processes, and that more mental processes are required when responding to negatively worded items (Bassili & Scott, 1996; Chessa & Holleman, 2007; Kamoen, Holleman, Mak, Sanders, & van den Bergh, 2011). In other words, not only the general latent trait, but also one specific factor for the wording effect are measured in negatively worded items. If the wording effect is not considered in data analysis, the essential assumption of local independence would be violated because the residuals between negatively worded items are correlated.

Global person dependence (GPD) is defined as the general similarity among people regarding measured proficiency on a test. It is convincing that, to some extent, people within the same cluster (or subgroup) may show more similar testing behavior than do people in different clusters. For example, pupils within a school have more similar academic ability than do those from different schools (Opdenakker, Van Damme, De Fraine, Van Landeghem, & Onghena, 2002); family members may provide similar perspectives when they are

recruited in the same survey (Deal, 1995; Jager, Bornstein, Putnick, & Hendricks, 2012); and citizens who live in the same geographical area have a more homogeneous incidence of a specific disease than do citizens in other areas (Langford, Leyland, Rasbash, & Goldstein, 1999). For some decades, researchers have attempted to deal with person-clustering structures by means of multilevel models (Bock, 1989; Burstein, 1980; Fox, 2005; Langford et al., 1999) so that GPD can be adequately quantified. The idea of multilevel modeling will be briefly reviewed in the next chapter.

Finally, LPD is defined as residual dependence among persons. In practice, it is often the case that, based on respondents' proficiency levels, local dependencies exist among persons' residuals. An example can be found in the context of ability tests, when response residuals for students within a cluster are not independent, i.e., apart from measured academic ability and measured item properties, an extra factor functioning among these students in their responses. For instance, answer copying (Belov, 2011; Cizek, 1999), which is very common on ability tests, entails a group of examinees showing illegitimate similarities in their response patterns. Test tempering, which is when students' responses are changed by teachers or invigilators after tests are completed, is another method of cheating on ability tests (Wollack, Cohen, & Eckerly, 2015). LPD also can be found when students in a class are taught to use a specific testing skill to respond to specific items (Eberbach & Crowley, 2009). In addition, a phenomenon that is well known in consumer research is when respondents in different countries exhibit substantial cross-national differences in the parameter estimates of items measuring consumer attitudes (de Jong, Steenkamp, & Fox, 2007; Steenkamp & Baumgartner, 1998), implying the possibility of LPD. In a national-level health census, the "neighborhood effect" was found, concerning the idea that neighborhood-level characteristics are associated with the occurrence of major mental disorders (Menezes, Georgiades, & Boyle, 2011). More examples are provided in the following discussion. It should be noted that multilevel

modeling assumes that observed responses are locally independent. When LPD exists behind observed responses, the local independence assumption is violated, and multilevel models will not function normally.

1.3 Importance

The main purpose of this study is to develop a new class of IRT models that incorporates the idea of random items into existing multilevel IRT models to account for LPD and to illustrate how these models can be carried out through Markov chain Monte Carlo (MCMC) estimation by using the freeware application WinBUGS (Spiegelhalter, Thomas, Best, & Lunn, 2007). The proposed approach has three substantial advantages. First, it recognizes the conceptual difference between GPD and LPD. In the new approach, GPD and LPD are exclusively formulated in different parameterizations so that the two kinds of dependence are no longer confounded. Second, it helps purify measurement scales because the contamination of potential LPD is sufficiently removed. Finally, the purified scale can improve the precision of person measures and ensure unbiased statistical inferences.

The specific research questions to be answered in this study are:

1. How does one develop a series of IRT models to consider GPD and LPD jointly in different testing scenarios?
2. Can the parameters in the proposed models be recovered by WinBUGS? What are the crucial factors in the parameter recovery?
3. What are the consequences of fitting standard multilevel IRT models when LPD is substantial, but ignored?
4. Except for the IRT modeling to LPD, how effectiveness is the performance of the conventional approach in detecting LPD?
5. How applicable are these methods to empirical data?

1.4 Overview of Chapters

The remainder of the dissertation is organized as follows. Relevant IRT models are reviewed in Chapter 2. In Chapter 3, an introduction to the new class of IRT models for clustered samples is presented, along with a non-IRT approach for detecting LPD. In Chapter 4, the designs for a series of simulations are presented to evaluate the parameter recovery of the new models and the consequences of ignoring LPD by fitting conventional models under various conditions. The applicability of HCA is also examined. In Chapter 5, four empirical examples are provided for illustration purposes. The presentation of these examples is arranged in line with the complexity of the implemented models. Finally, in Chapter 6, I summarize the main findings and conclude the study by presenting a discussion of the results, along with directions for future research.

Chapter 2: Literature Review

2.1 Review of standard IRT models

IRT models were developed to analyze categorical responses (Embretson & Reise, 2000; Lord, 1980). For example, the famous Rasch model (Rasch, 1960) defines the log-odds of the two probabilities for a dichotomous response as:

$$\log\left(\frac{P_{ni1}}{P_{ni0}}\right) = \theta_n - \delta_i, \tag{2}$$

in which P_{tni0} and P_{tni1} stand for the probabilities of scoring 1 and 0, respectively, on item i (i = 1,…, I) for person n (n = 1,…, N); θ_n is the latent ability of person n; and δ_i is the difficulty of item i. The dichotomous Rasch model is also called the one-parameter logistic model because only a single parameter (i.e., item difficulty) is specified for each item. Likewise, the two- and three-parameter logistic models (Birnbaum, 1968) are proposed by adding slope and asymptotic parameters. The slope parameter is commonly interpreted as the item discrimination, whereas the asymptotic parameter is viewed as the guessing parameter, especially for multiple-choice items.

Many IRT models have been developed to analyze polytomous responses. Let there be an item scored as 0, 1, …, J. For example, the partial credit model (PCM; Masters, 1982), which is one of the Rasch family models for categorical data, can be expressed as:

$$\log\left[\frac{P_{nij}}{P_{ni(j-1)}}\right] = \theta_n - \delta_{ij}, \tag{3}$$

in which P_{nij} and $P_{ni(j-1)}$ are the probabilities of endorsing options j and $j-1$ on item i (i = 1,…, I) for person n (n = 1,…, N); θ_n is the measured latent trait for person n; and δ_{ij} is the j-th threshold parameter for item i. When the same scoring rubric is applied to all items, one can impose the constraint that the differences between two adjacent thresholds across items are equal:

$$\log\left[\frac{P_{nij}}{P_{ni(j-1)}}\right] = \theta_n - (\delta_i + \tau_j), \quad (4)$$

in which τ_j is the j-th deviation from the δ_i. Equation 4, thus, becomes the rating-scale model (RSM; Andrich, 1978). Similarly, if the PCM is generalized by incorporating the slope parameter, then the generalized partial credit model is formed (Muraki, 1992).

In the aforementioned models, the analyzed samples are treated as unique individuals and are deemed independent of each other, i.e., these models are unable to indicate the similarity between and within clusters. Certainly, one can investigate how similar the latent traits among a group of persons are by means of person estimates, but the derived values may be attenuated because the measurement errors of person estimates are neglected in the subsequent analysis (Mislevy, 1991).

2.2 Multilevel modeling

In large-scale educational and psychological testing programs, test-takers often have a multilevel structure. For example, in the Program for International Student Assessment (PISA; Organization for Economic Cooperation and Development, 2014), approximately 150 schools are first randomly selected from a country, and approximately 40 students are then randomly sampled from each sampled school. Such a multiple-stage sampling creates a multilevel data structure.

Multilevel modeling is a generalization of regression models (Stephen & Anthony, 2002). For example, let g (g = 1, …, G) index the groups (e.g., schools) and Y_{ng} be the academic ability of person n in group g. At Level 1, Y_{ng} can be regressed on a set of Level 1 predictors x_1, …, x_Q, such as gender, socioeconomic status, and IQ. At Level 2, the regression parameters at Level 1 can be further regressed on group predictors w_1, …, w_S, such as school type and school size:

$$\text{Level 1:}\quad Y_{ng} = \eta_{0g} + \sum_{q=1}^{Q} \eta_{qg} x_{qng} + e_{ng}, \quad (5)$$

$$\text{Level 2:}\quad \eta_{0g} = \gamma_{00} + \sum_{s=1}^{S} \gamma_{0s} w_{sg} + u_{0g}\,, \tag{6}$$

$$\eta_{1g} = \gamma_{10} + \sum_{s=1}^{S} \gamma_{1s} w_{sg} + u_{1g}\,, \tag{7}$$

$$\vdots$$

$$\eta_{Qg} = \gamma_{Q0} + \sum_{s=1}^{S} \gamma_{Qs} w_{sg} + u_{Qg}\,. \tag{8}$$

It is assumed that $e_{ng} \sim N(0, \sigma_e^2)$, and $\mathbf{u}_g \sim MVN(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{u}})$. If data have more than two levels (e.g., schools are randomly selected from school districts), equations 6–8 can continue for more levels.

In recent years, IRT-based multilevel models have been developed (Fox, 2005; Fox & Glas, 2001; Kamata, 2001; Katama & Vaughn, 2011; Maier, 2001) so that item parameters, as well as regressed coefficients, can be estimated jointly in one computer run. For example, when considering a simple, two-level structure (without predictors at Level 1 or 2) in the Rasch model, the multilevel Rasch model (MRM) becomes:

$$\log\left(\frac{P_{ngi1}}{P_{ngi0}}\right) = \theta_{ng} - \delta_i \equiv (\eta_g + \varepsilon_n) - \delta_i\,, \tag{9}$$

in which θ_{ng} is the latent trait of person n within group g; η_g is the mean of θ_{ng} for group g; and ε_n is the regressed residual following a normal distribution with a mean of zero and variance of $\sigma_\varepsilon^{2\backprime}$. If the dataset includes many groups, say, more than 30, it may be reasonable to assume that η_g follows a normal distribution with a mean of μ_η and a variance of σ_η^2. Because the MRM is a random-intercept model with a logit link (Roudenbush & Bryk, 2002), the intra-class coefficient (ICC) can be referred to as the effect size of GPD:

$$ICC = \frac{\sigma_\eta^2}{\sigma_\eta^2 + \sigma_\varepsilon^2}\,. \tag{10}$$

According to Hox (2010), the values of .05, .10, and .15 could be the guidelines for small, medium, and large effect sizes, respectively, in judging the extent of ICC. When Level 1 or Level 2 covariates are available, the latent variable θ_{ng} can be further regressed, such as the decompositions in equations 5–8.

Although multilevel models successfully take multilevel structures into account, this does not imply that the local independence assumption is no longer required. Conditional on θ_{ng}, the responses between persons are still assumed to be locally independent, i.e., multilevel modeling is a useful technique for quantifying GPD, but fails to take LPD into consideration.

2.3 Modern IRT models for LPD

Few IRT models have been developed specifically to test the LPD assumption among subjects' responses. One of the IRT models is the response dependence of subjects model (RDSM; Cristante & Robusto, 1999; Robusto & Cristante, 2010), which belongs to the family of Rasch models and can quantify dependence within small groups. Unlike common IRT models, the analyzed unit in the RDSM is a well-defined cluster (e.g., a family), rather than individual respondents. The RDSM is essentially built upon the binomial model and takes the form of Equation 11:

$$\log\left[\frac{P_{gix}}{P_{gi(x-1)}}\right] = \beta_g - \left[\kappa_i + \psi_g(2x - N_g - 1)\right], \tag{11}$$

in which P_{gix} and $P_{gi(x-1)}$ are the probabilities of getting total scores x and $x - 1$ on item i for cluster g ($g = 1,\ldots, G$); N_g is the size of cluster g; β_g is the location on the measured scale of cluster g; κ_i is the location parameter of item i; and ψ_g is the dependence parameter of cluster g. The RDSM treats a total of N_g respondents belonging to the same cluster as replications of the same event, and it assumes that these respondents have an equal probability of getting a score on an item. As noted, the chi-square statistic can be applied on a contingency table, in which the number of items chosen by each person is compared with the other persons'

decisions to test the adequacy of equal probability. The value of ψ_g determines the shape of the binomial distribution. When $\psi_g = 0$, the distribution is uniform; when $\psi_g > 0$, the distribution is unimodal; and when $\psi_g < 0$, the distribution is concave. Therefore, the occurrence of person dependence can be judged according to the interaction of the value of ψ_g by equal probability. Cristante and Robusto (1999) illustrated that when ψ_g is less than a critical value, it can be concluded that the responses for members within cluster *g* are dependent, whereas when ψ_g is larger than the critical value, both dependence and independence are possible because independence can be made on the basis of unequal probability.

Although the RDSM can isolate the influence of person dependence successfully, it has three major limitations in its application. First, the items must be dichotomous, which is a serious limitation because most survey inventories adopt polytomous items. In other words, if the survey items are designed in a Likert-type scale structure, the RDSM becomes inapplicable. Second, although one can test the assumption of equal probability in the RDSM, an estimate of the latent trait for each respondent is not available. Such a strategy may cause one to question whether it makes sense to combine individual scores (Ganong, 2003). Finally, the RDSM accounts for general person dependence at the test level, but it ignores the reality that the extent of dependence may vary across items.

A multilevel model for dual local dependence (Jiao, Kamata, Wang, & Jin, 2012) was developed to account for item and person clustering simultaneously. The model is an exact conjunction of multilevel IRT models and testlet models:

$$\log\left(\frac{P_{ngid1}}{P_{ngid0}}\right) = \theta_n + \theta_g - \delta_i + o_{nd(i)}, \tag{12}$$

in which P_{ngid1} and P_{ngid0} are the probabilities of scoring 1 and 0 for person *n* on item *i* within testlet *d*; θ_n is the person-specific ability for person *n*; θ_g is the group-specific ability for

cluster g; δ_j is the difficulty for item i; and $o_{nd(i)}$ is the random effect of person n on item i within testlet d. Accordingly, the variance of $o_{nd(i)}$ stands for the magnitude of item dependence, whereas the variance of θ_g represents the magnitude of person dependence. Equation 12 can be easily generalized to two- or three-parameter models and more complex models for polytomous responses.

In comparing Equations 9 and 12, one can conclude that the multilevel model proposed by Jiao et al. (2012) is an extension of Equation 9, which considers LID among items in a testlet additionally. Based on the illustration in multilevel modeling, it suggests that Jiao et al. (2012) assumed local independence between persons, i.e., they considered GPD rather than LPD.

2.4 Differential item functioning models

The occurrence of LPD is relevant to the issue of measurement non-invariance. When LPD is observed, it refers to an unfair situation in which participants with the same level of a latent trait, but from different clusters, have an unequal probability of endorsing an item. Such an unfair situation is usually interpreted as differential item functioning (DIF; Holland & Wainer, 1993) in the literature because the item-characteristic curves vary across groups. In recent decades, many statistical techniques have been developed for DIF assessment, and they can be classified roughly into IRT-based and non-IRT approaches (Magis, Béland, Tuerlinckx, & De Boeck, 2010). IRT-based approaches, such as the likelihood ratio test (Cohen, Kim, & Wollack, 1996), Lord's chi-square test (Lord, 1980), Raju's signed area method (Raju, 1988, 1990), and multiple indicators and multiple causes method (Finch, 2005; Wang & Shih, 2010; Wang, Shih, & Yang, 2009), are fit to data, and statistical tests are implemented to compare item parameters (or derived item characteristic curves) from different groups. Consequently, an item is flagged as DIF if a difference of item parameters between groups is significant. Non-IRT approaches in DIF assessment include the

Mantel-Haenszel method (Holland & Thayer, 1988; Mantel & Haenszel, 1959), logistic regression (Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990), the delta method (Angoff & Ford, 1973), standardization (Dorans & Kulick, 1986), and SIBTEST (Shealy & Stout, 1993), among others. These procedures do not have any requirements for specific forms of item-response functions and large sample sizes; therefore, they are computationally simple.

When modeling DIF in the dichotomous Rasch model, the log-odds of the two probabilities of item *i* for person *n* within group *g* are expressed as:

$$\log\left(\frac{P_{ngi1}}{P_{ngi0}}\right) = \theta_{ng} - \delta_i + \varphi_{ig}, \tag{13}$$

in which φ_{ig} stands for the group-specific advantage/disadvantage for group *g*, and the sum of φ_{i1}, φ_{i2},…, φ_{iG} is constrained at zero for identification. Because the influence of DIF formulated in Equation 13 is uniform across the latent continuum within a group, the differential function on item difficulty is also called uniform DIF. Usually, respondents are classified into a focal group (F) and a reference group (R) in most DIF studies. When investigating whether item *i* is fair toward the two groups, the null hypothesis of $\varphi_{iF} = \varphi_{iR} = 0$ is set. The null hypothesis is rejected, and item *i* is flagged as DIF when the DIF size of $|\varphi_{iF} - \varphi_{iR}|$ is significantly larger than the critical value; otherwise, item *i* is temporarily deemed DIF-free. Linacre (2016) provided a guideline, similar to the Educational Testing Service (ETS) DIF category, for judging the effect size of DIF in the Rasch models. Category A (DIF < 0.43) comprises items with negligible or no DIF, Category B (DIF ≥ 0.43) comprises items with slight-to-moderate DIF, and Category C (DIF ≥ 0.64) comprises items with moderate-to-large DIF.

There are three major approaches to DIF detection. The first is the multiple-group comparison approach, in which one reference group is identified and compared with multiple

focal groups on a DIF statistic for each item (Finch, 2016; Kim, Cohen, & Park, 1995; Magis, Raîche, Béland, & Gérard, 2011; Penfield, 2001). The advantage of the multiple group comparison approach is its conceptual simplicity. The selected univariate statistic for the DIF assessment between two groups could be generalized to a multivariate statistic for conducting multiple-group comparisons simultaneously. When two or three grouping factors (e.g., male and female; Asian and non-Asian) are available, traditionally, one could conduct multiple DIF analyses by considering one grouping factor at a time, or one could conduct one analysis by combining all grouping factors into a pseudo grouping factor (e.g., Asian male, Asian female, non-Asian male, and non-Asian female). The second is the factorial analysis of variance (ANOVA) approach (Jin, Chen, & Wang, 2015; W.-C. Wang, 2000), in which multiple grouping factors are jointly included in an analysis to decompose their influences into main effects and interactions. It outperforms the two traditional analyses, especially when grouping factors exhibit interactions. Noticeably, when large numbers of groups are considered, the third approach, namely the random DIF approach, is the most efficient (de Jong & Steenkamp, 2010; de Jong, Steenkamp, & Fox, 2007), i.e., DIF across groups is assumed to follow a normal distribution:

$$\varphi_{ig} \sim N(0, \sigma^2_{\varphi_i}) . \tag{14}$$

Thus, the variance of φ_i indicates the magnitude of DIF for item *i*.

One can realize the random DIF approach as a type of cross-random effect model. As commented by De Boeck (2008), not only the person parameters, but also the item parameters can be treated as random effects. The term *random item parameter* actually functions in two forms. One is *random across items*, referring to when using a random distribution is feasible to account for the characteristics of items gathered in a pool. The other one is *random across persons within items*, referring to when the interaction between person and item contributes to the randomness of item parameters, i.e., the random DIF approach is

aligned with the concept of *random across person groups/clusters within items*.

Chapter 3: Methodologies

3.1 Model development

Adding a random variable into IRT models is a generic method for representing LID (Jiao et al., 2012; Wainer et al., 2007; Wang et al., 2015; Wang & Wilson, 2005). Adopting this strategy, LPD is modeled by adding a random variable for persons within a cluster on each item. For example, the MRM (i.e., Equation 9) can be generalized as:

$$\log\left(\frac{P_{ngi1}}{P_{ngi0}}\right) = \theta_{ng} - \delta_i + \xi_{ig}, \tag{15}$$

in which ξ_{ig} represents the LPD among persons within cluster g on item i, and the other variables have been defined as above. Because the fixed-effect approach is inapplicable when the number of clusters increases, it is practical to assume $\xi_{ig} \sim N(0, \sigma^2_{\xi_i})$. In addition, when considering a two-level structure, θ_{ng}, the latent trait of person n within group g, can be regressed by cluster-level and person-level predictors:

$$\theta_{ng} = \mu + \sum_{v=1}^{V} \gamma_v X_{vg} + \upsilon_g + \sum_{u=1}^{U} \eta_u W_{ung} + \varepsilon_{ng}, \tag{16}$$

in which $\boldsymbol{X}$ and $\boldsymbol{W}$ refer to the vectors for cluster-level and person-level predictors; $\boldsymbol{\gamma}$ and $\boldsymbol{\eta}$ are the corresponding regression coefficients; $\boldsymbol{\upsilon}$ and $\boldsymbol{\varepsilon}$ are the vectors for regressed residuals; and μ is the grand mean for θ_{ng} and is constrained at zero for identification. Equations 15 and 16 are named the multilevel Rasch model for clustered samples (MRM-CS), which takes the GPD (i.e., the multilevel structure on θ) and LPD (i.e., ξ_{ig}) jointly into account. For each item, the target latent trait θ and one person dependence factor ξ are measured, indicating that a total of $I + 1$ random variables are included in the measurement model. In addition, these random variables are assumed to be mutually independent. Because each ξ parameter is measured by a single item, precise estimates for individual clusters are not possible, but a precise estimate of $\sigma^2_{\xi_i}$ is achievable with a sufficiently large sample. The parameter $\sigma^2_{\xi_i}$

characterizes the magnitude of LPD in item *i*: The larger the value of $\sigma^2_{\xi_i}$, the larger the LPD for that item will be. When $\sigma^2_{\xi_i}$ is zero for every item of a scale, the MRM-CS simplifies to the MRM. Statistically, the MRM-CS can be viewed as a composition of three elements: the Rasch model, the multilevel structure on the latent trait, and the random DIF model.

If LPD exists ($\sigma^2_{\xi_i} \neq 0$), but is ignored by fitting a standard model, a shrunken scale would be obtained due to the nonlinear relationship between the probability and logit function. One can understand the causality through the following illustration. Let's say there's a dichotomous item following the MRM-CS with a difficulty of 0. Conditional on the five ability levels of -2, -1, 0, 1, and 2, the expected probabilities at the five ability levels are .12, .27, .50, .73, and .88, respectively. Suppose there is a large LPD $\sigma^2_{\xi_i} = 1$ on that item: The marginal probabilities at the five ability levels become .16, .30, .50, .70, and .84, respectively. When projecting these marginal probabilities onto the item characteristic curve in the Rasch model, the abilities of -1.66, -0.85, 0, 0.85, and 1.66, respectively, can be derived, suggesting that ignoring LPD would lead to a shrunken scale. Consequently, the estimates of σ^2_{η} and σ^2_{ε} would be smaller than their true values.

It is straightforward to apply the strategy of modeling GPD and LPD to polytomous responses. For example, the PCM (i.e., Equation 3) can be extended as:

$$\log\left[\frac{P_{ngij}}{P_{ngi(j-1)}}\right] = \theta_{ng} - \delta_{ij} + \xi_{ig}. \qquad (17)$$

Thus, Equation 17 is referred to as the multilevel partial credit model for clustered samples (denoted as MPCM-CS). When necessary, it is feasible to consider LPD together with different kinds of LID in the same model, as well as the dual dependence model proposed by Jiao et al. (2012):

$$\log\left[\frac{P_{ngij}}{P_{ngi(j-1)}}\right] = \theta_{ng} - \delta_{ij} + \xi_{ig} + o_{nd(i)}. \quad (18)$$

The paired sample design, a special case of person clustering, is widely applied in psychological, biological, and medical experiments. In paired samples, there are two members -- e.g., a husband and wife -- in each cluster. Two people, as a pair, can be categorized into a focal unit and a reference unit regarding a demographic variable. Equation 16 then can be rewritten for paired samples:

$$\text{For focal unit}: \log\left[\frac{P_{gij}}{P_{gi(j-1)}}\right]_{\text{F}} = \theta_g + \theta_{g\text{F}} - \delta_{ij} + \xi_{ig}, \quad (19)$$

$$\text{For reference unit}: \log\left[\frac{P_{gij}}{P_{gi(j-1)}}\right]_{\text{R}} = \theta_g + \theta_{g\text{R}} - \delta_{ij} + \xi_{ig}, \quad (20)$$

in which θ_g is the mean ability of pair g, and $\theta_{g\text{F}}$ and $\theta_{g\text{R}}$ are the person-specific deviations from θ_g for focal and reference units, respectively. Thus, it is assumed that $\theta_g \sim N(\mu_g, \sigma_g^2)$, $\theta_{g\text{F}} \sim N(0, \sigma_{\text{F}}^2)$, and $\theta_{g\text{R}} \sim N(0, \sigma_{\text{R}}^2)$. A constraint of $\sigma_{\text{F}}^2 = \sigma_{\text{R}}^2$ is viable, depending on the context. Instead of the multilevel modeling approach, Jin and Wang (2016) recently proposed a bivariate normal distribution to describe the two latent traits of paired samples. The multilevel parameterization, such as Equations 19 and 20, outperforms the bivariate normal distribution approach, especially when cluster-level covariates are included to explain the impacts on θ.

The concept of clustered samples can go beyond the virtual person clustering structure. In repeated measures design, for example, an individual becomes a cluster, and observations at different time points are subjects within individuals (Hox, 2010; Hox & Roberts, 2011). Particularly, when a set of common items elicits responses from the same people more than once, it may result in memory effects, a type of LPD among residuals across time points within a person (Olsbjerg & Christensen, 2014). Accordingly, the MRM-CS and MPCM-CS

can be directly applied to this situation.

As stated in Chapter 2.4, LPD can be viewed as random DIF. As summarized by Cho, Suh, and Lee (2016), there are mainly four methods for dealing with the presence of DIF items in the literature: (a) removing DIF items, (b) ignoring DIF items, (c) calibrating item parameters for different groups separately, and (d) modeling DIF. Particularly, approach (a) is actually identical to no treatment by fitting a standard model, and approach (d) is close to the proposed models for dealing with LPD. Approach (c) seems inapplicable because the number of clusters could be very huge. There is an interpretive difference between the viewpoints of DIF and LPD. The purpose of DIF detection is to shed light on the extent to which unfair items are included in a test. By definition, a DIF-free item can be presumed only if its item characteristic curve is uniform across groups; otherwise, that item would be labeled a DIF item. Consequently, the interpretation of DIF becomes a dichotomy, in that an item is eventually classified as either DIF or non-DIF. Researchers are skeptical when they see too many DIF items in a test because the occurrence of DIF items implies test unfairness. In contrast, the idea of dichotomy is not embraced in the interpretation of LPD. Sometimes, LPD is anticipated. For instance, when couples or family members are recruited in a survey, both their similarity and dissimilarity on thoughts and attitudes are equally of interest. On the other hand, technically, DIF could result from a more complex person-item interaction, whereas a simpler person-item interaction is deliberated in the proposed approach. The definition becomes more complex for polytomous items. Consider a five-point Likert-type item in which four threshold parameters are modeled under the PCM. If one or more of the four threshold parameters interact with the grouping variable, a nonuniform item characteristic curve is achieved. The complexity in the combinations of thresholds increases the difficulty of explanation. Comparatively, a simpler mechanism was implemented when modeling LPD. As illustrated in Equations 15 and 17, a single parameter (i.e., ξ_{ig}) is referred

to as the unit of translation of the whole item characteristic curve for cluster *g* (related to δ_i), and the value of $\sigma^2_{\xi_i}$ reflects the magnitude of LPD directly.

The number of clusters and intra-cluster sample sizes would influence the precision of parameter estimation. Literally, a cluster should include more than one subject. Jin and Wang (2015) demonstrated that LPD parameters can be accurately estimated in paired-sample data, i.e., the intra-cluster sample size may be not a critical influence on the estimation of $\sigma^2_{\xi_i}$, as long as the number of clusters is large enough. The influences of cluster numbers and intra-cluster sample sizes on the precision of LPD parameters will be examined in Chapter 4.2.

3.2 Extensions to multiple scales

The extension of the LPD model includes many axes. In the aforementioned models, only a target latent trait θ is measured. In practice, however, a test may consist of multiple scales, in which a different latent trait is measured on each subscale. For example, the International English Language Testing System (IELTS) measures four kinds of proficiency: listening, speaking, reading, and writing. It has been noted that utilizing the multidimensional approach is more efficient in reporting the correlation between latent traits and test reliabilities than the consecutive uni-dimensional approach, in which multiple scales are analyzed separately, one at a time (Adams, Wilson, & Wang, 1997; Briggs & Wilson, 2003). To analyze multiple scales jointly, Equations 15 and 17 can be extended as

$$\log\left(\frac{P_{ndgi1}}{P_{ndgi0}}\right) = \theta_{ndg} - \delta_i + \xi_{ig}\,, \tag{21}$$

$$\log\left[\frac{P_{ndgij}}{P_{ndgi(j-1)}}\right] = \theta_{ndg} - \delta_{ij} + \xi_{ig}\,. \tag{22}$$

in which subscript *d* (*d* = 1,…, *D*) is the index of scale. Finally, vector $\boldsymbol{\theta} = [\theta_1, ..., \theta_d]'$ contains *D* elements, and they can be estimated simultaneously in the joint model. Equation

21 is referred to as the multidimensional multilevel Rasch model for clustered samples (MDMRM-CS), and Equation 22 is referred to as the multidimensional multilevel partial credit model for clustered samples (MDMPCM-CS). Note that these multidimensional models are not limited to between-item multidimensional tests but can be applied to tests with a within-item multidimensional structure (i.e., an item measures more than one latent trait).

Furthermore, considering the homogeneity of a test and the heterogeneity among subtests, it is feasible and manageable to build a hierarchical structure that treats the scores derived from subtests as subordinate elements under a higher-order score, built on the idea of higher-order factor analysis (Matin & Adkins, 1954). Recently, several hierarchical IRT models for hierarchical latent traits have been developed (de la Torre & Hong, 2010; de la Torre & Song, 2009; Sheng & Wikle, 2008; Huang & Wang, 2013; Huang, Wang, Chen, & Su, 2013). In their models, the first-order latent trait is assumed to be a weighted function of the second-order latent trait:

$$\theta_{nd}^{(1)} = \lambda_d^{(2)}\theta_n^{(2)} + \upsilon_{nd}^{(1)}, \tag{23}$$

in which $\theta_{nd}^{(1)}$ is a first-order latent trait measured in the *d*th subtest for respondent *n*; $\theta_n^{(2)}$ is the second-order latent trait; $\lambda_d^{(2)}$ is a regression weight of the second-order latent trait on the *d*th first-order latent trait; and $\upsilon_d^{(1)}$ is the residual and is assumed to be normally distributed. Because hierarchical latent traits and the person-clustering structure can occur simultaneously, a general IRT model – which considers GPD, LPD, and hierarchical latent traits altogether – is applicable. For example, the higher-order, partial-credit model for clustered samples (HOPCM-CS) can be reformulated as:

$$\log\left[\frac{P_{ngij}}{P_{ngi(j-1)}}\right] = \theta_{ndg}^{(1)} - \delta_{ij} + \xi_{ig} = \left(\lambda_d^{(2)}\theta_{ng}^{(2)} + \upsilon_{ndg}^{(1)}\right) - \delta_{ij} + \xi_{ig}\,. \tag{24}$$

Following this strategy, a more complex model with more orders can be formed when necessary. Statistically, higher-order models can be linked to bi-factor models (Brown, 2015).

Thus, Equation 24 can be reparameterized to a bi-factor model:

$$\log\left[\frac{P_{ngij}}{P_{ngi(j-1)}}\right] = \left(\theta'_{ng} + \upsilon'_{ndg}\right) - \delta_{ij} + \xi_{ig}, \tag{25}$$

in which θ'_{ng} is the measured latent trait across subtests, and υ'_{ndg} is the domain-specific nuisance of subtest *d*. The parameters in Equations 24 and 25 are comparable after the standardization of θ parameters.

3.3 Extensions to multifaceted data

The realization of LPD is not necessarily limited to item-cluster interaction, but could be generalized to other situations. For example, multifaceted data (e.g., test-takers' responses to items marked by raters) are very common in the human sciences, and the many-faceted Rasch model (MFRM; Linacre, 1989) is widely used in practice, partly due to its simplicity (Basturk, 2008; Congdon & MeQueen, 2000; Eckes, 2005, 2008; Engelhard, 1994, 1996; Myford & Wolfe, 2003, 2004; Schaefer, 2008). In the MFRM, an individual element within a facet is assigned a parameter to indicate its influence on item responses. In the three-faceted data of rater, criterion, and ratee, for example, there are three kinds of parameters: an item's difficulty, a ratee's proficiency, and a rater's severity. The three-faceted Rasch model is expressed as:

$$\log\left[\frac{P_{ngijk}}{P_{ngi(j-1)k}}\right] = \theta_{ng} - \delta_{ij} - \iota_k, \tag{26}$$

in which θ_{ng} is the proficiency of ratee *n* within group *g*; δ_{ij} is the *j*th threshold difficulty of criterion *i*; and ι_k is the severity of rater *k*. When LPD is formulated as the rater-ratee interaction, it is often referred to as differential rater functioning (DRF) (Du, Wright, & Brown, 1996; Engelhard, 2008), which means a rater may exhibit different severities for rates within different clusters. Thus, the many-faceted model for clustered samples (MFRM-CS) can be expressed as:

$$\log\left[\frac{P_{ngijk}}{P_{ngi(j-1)k}}\right] = \theta_{ng} - \delta_{ij} - \iota_{k} + \xi_{kg}, \tag{27}$$

in which ξ_{kg} represents the LPD toward rater k (k = 1,…, K), and others are defined as stated above. The proposed methodology definitely can be applied to data with more than three facets.

Although the concept of LPD can be generalized as ratee-rater interaction when ratees are grouped, this is not the only case. Because raters are human beings, it is very likely that the ratings among raters are not necessarily independent, especially when raters are allowed to communicate with each other before providing ratings on a criterion (Wang, Su, & Qiu, 2014). To model LPD among raters, the many-faceted model can be extended to:

$$\log\left[\frac{P_{nijkm}}{P_{ng(j-1)km}}\right] = \theta_{n} - \delta_{ij} - \iota_{km} + \xi_{im}, \tag{28}$$

in which ι_{km} is the severity of rater k within rater group m (m = 1,…, M), and ξ_{im} represents LPD among raters toward item i.

3.4 More model extensions

More extensions of IRT models for LPD are possible. For example, within the two-parameter IRT framework, a slope parameter is incorporated. Hence, the MRM-CS and MPCM-CS can be generalized to:

$$\log\left(\frac{P_{ndgi1}}{P_{ndgi0}}\right) = \alpha_{i}\left(\theta_{ng} - \delta_{i} + \xi_{ig}\right), \tag{29}$$

$$\log\left[\frac{P_{ndgij}}{P_{ndgi(j-1)}}\right] = \alpha_{i}\left(\theta_{ndg} - \delta_{ij} + \xi_{ig}\right), \tag{30}$$

in which α_i is the slope parameter and stands for the discrimination power of item i with respect to measured latent trait θ. Equations 29 and 30 can be referred to as the multilevel two-parameter logistic model for clustered samples (M2PLM-CS) and the multilevel

generalized partial credit model for clustered samples (MGPCM-CS), respectively. Note that the relationships among the parameters in both models become nonlinear due to the products of the parameters between α_i and θ_n. When $\alpha_i = 1$ for all items, the M2PLM-CS simplifies to the MRM-CS, and the MGPCM-CS simplifies to the MPCM-CS.

The proposed models can be generalized to mixture models. Although one can include any observed group membership in a multilevel model, sometimes momentous group memberships are unknown or latent, i.e., except for manifest grouping variables, respondents can be grouped into clusters based on unobserved variables as well. For example, it is believed that unmotivated respondents usually answer items casually, without full consideration of the item content (Johnson, 2005; Meade & Craig, 2012; Nichols, Greene, & Schmolck, 1989), but respondents' exertion is latent and unidentified in the dataset. Thus, it is important to include respondents' latent group membership in multilevel models. When the latent group membership is considered in the MRM-CS, for example, the generalized model is formulated as:

$$\log\left(\frac{P_{ngli1}}{P_{ngli0}}\right) = \theta_{ngl} - \delta_{il} + \xi_{igl}\,, \tag{31}$$

in which l ($l = 1, \ldots, L$) is an index for person-level latent classes; θ_{ngl} is the latent trait of person n in manifest cluster g and latent class l; δ_{il} is the difficulty of item i for latent class l; and ξ_{igl} is the LPD among persons within cluster g in latent class l on item i. Moreover, based on the multilevel mixture IRT model proposed by Cho and Cohen (2010), latent classes are not limited to the person level and can exist at the cluster level.

3.5 Model parameter estimation

Marginal maximum likelihood estimation is widely used in parameter calibration for IRT models (Johnson, 2007; Tuerlinckx et al., 2004). The technique simply assumes that individuals are sampled from a large distribution so that the marginal probability of observed

responses can be obtained by integrating the random variables from the likelihood function. However, high-dimensional integration in the proposed models for LPD became a problem, making marginal maximum likelihood estimation unfeasible.

Alternatively, I have adopted the Bayesian approach with MCMC estimation, which is very efficient for achieving a high dimensional integral. In Bayesian estimation, a statistical model and prior distributions of model parameters are specified to yield a joint posterior distribution. MCMC methods provide alternative and simple ways to simulate the joint posterior distribution of the unknown quantities and obtain simulation-based estimates of the posterior parameters of interest. For example, the likelihood function for the MPCM-CS is expressed as:

$$L(\mathbf{Y} \mid \boldsymbol{\varepsilon}, \boldsymbol{\upsilon}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\xi}, \boldsymbol{\delta}) = \prod_{g}^{G} \prod_{n=1}^{N_g} \prod_{i=1}^{I} \prod_{j=1}^{J} (P_{ngij})^{y_{ngij}} \ . \tag{32}$$

Notably, ε_{ng}, υ_g, and ξ_{ig} are random variables following different normal distributions independently:

$$\varepsilon_{ng} \sim N(0, \sigma_{\varepsilon}^2),$$

$$\upsilon_g \sim N(0, \sigma_{\upsilon}^2), \text{ and}$$

$$\xi_{ig} \sim N(0, \sigma_{\xi_i}^2).$$

Through the above specifications, σ_{ε}^2, σ_{υ}^2, and $\sigma_{\xi_i}^2$ are estimated instead, and the prior of $\Gamma(0.1, 0.1)$ is implemented for $1/\sigma_{\varepsilon}^2$, $1/\sigma_{\upsilon}^2$, and $1/\sigma_{\xi_i}^2$ in WinBUGS. Moreover, the prior of $N(0, 10)$ is implemented for each of the $\boldsymbol{\gamma}$, $\boldsymbol{\eta}$, and $\boldsymbol{\delta}$ parameters. Less-informative priors could be applied unless the sample size is limited. Finally, the joint posterior distribution of the estimated parameters can be derived:

$$\begin{aligned} p(\boldsymbol{\gamma}, \boldsymbol{\eta}, \sigma_{\varepsilon}^2, \sigma_{\upsilon}^2, \sigma_{\xi}^2, \boldsymbol{\delta} \mid \mathbf{Y}) \propto\ & L(\mathbf{Y} \mid \boldsymbol{\gamma}, \boldsymbol{\eta}, \sigma_{\varepsilon}^2, \sigma_{\upsilon}^2, \sigma_{\xi}^2, \boldsymbol{\delta}) \times \\ & p(\boldsymbol{\eta} \mid \boldsymbol{\gamma}, \sigma_{\varepsilon}^2, \sigma_{\upsilon}^2, \sigma_{\xi}^2) p(\boldsymbol{\gamma} \mid \sigma_{\upsilon}^2, \sigma_{\xi}^2) p(\sigma_{\varepsilon}^2) p(\sigma_{\upsilon}^2) p(\sigma_{\xi}^2) p(\boldsymbol{\delta}) \end{aligned} . \tag{33}$$

After sequential sampling, the posterior distribution of each parameter is formed. Its mean and standard deviation could be reported as the point estimate and the corresponding standard error of the parameters. The freeware application WinBUGS was used in this study. After a 5,000-iteration burn-in period, the subsequent 5,000 iterations with a thinning rate of 10 were used to compute the parameter estimates.

Bayesian model-data fit can be conducted by posterior predictive model checking (PPMC), which assesses the plausibility of posterior predictive replicated data against observed data and has the advantage of a strong theoretical basis and an intuitively appealing simplicity that can be applied to numerical evidence (Gelman, Meng, & Stern, 1996):

$$p = \Pr[S(\mathbf{Y}^{\text{rep}}) \geq S(\mathbf{Y}) \mid \mathrm{Y}], \tag{34}$$

in which S(·) denotes the used statistical index. In this study, the Bayesian chi-square test (Sinharay, 2005; Sinharay, Johnson, & Stern, 2006), which assesses the overall model-data fit, is chosen for detecting the systematic discrepancy between the observed and replicated data given the model parameters. An extreme *p*-value (close to zero or one) indicates model-data misfit.

Moreover, the Q_3 statistic (Yen, 1984) is adopted for assessing the extent of local dependence among item responses. When investigating the dependence between the residuals for two items, Q_3 is defined as:

$$Q_3 = cor\left(\mathbf{r}_i, \mathbf{r}_{i'}\right), \tag{35}$$

in which $\mathbf{r}_i$ and $\mathbf{r}_{i'}$ are the vectors of the residuals of items *i* and *i*', and Q_3 is the correlation between the two residual scores. When the assumption of local item independence is held, Yen (1993) demonstrated that the expected value of Q_3 is approximately -1/(*I* - 1) (*I* is the number of items). Chen and Wang (2007) further determined that the expected standard deviation is approximately $\sqrt{1/N-2}$ (*N* is the number of persons). When applying the Q_3 statistic to investigate the dependence between the residuals of two persons, the residual

matrix is transposed, and the computation of Q_3 becomes:

$$Q_3 = cor(\mathbf{h}_n, \mathbf{h}_{n'}), \tag{36}$$

in which $\mathbf{h}_n$ and $\mathbf{h}_{n'}$ are the vectors of the residuals of person n and n', and Q_3 is the correlation between the two vectors of residual scores. Accordingly, under the null hypothesis of the local independence of person residuals, Q_3 will be approximately normally distributed with mean -1/(N – 1) and standard deviation $\sqrt{1/T-2}$ when the sample size is large.

The Akaike information criterion (AIC), Bayesian information criterion (BIC), and deviance information criterion (DIC) are commonly used to compare the models (e.g., Hung, 2011; Li, Jiao, & Macready, 2016). Thus, these indices were selected for model comparison. However, Jin and Wang (2016) indicated that these indices are not sensitive and always favor the complicated model with a large number of random-effects parameters incorporated. To check the model-data fit for the standard and new models, the PPMC p-value and the Q_3 were mainly implemented in this study.

3.6 Detection of LPD by cluster analysis

Except for the LPD modeling approach, existing non-IRT approaches may be useful for detecting LPD. The occurrence of LPD suggested that, after fitting a standard model, the residuals of respondents who are nested within the same cluster are correlated. Consequently, the residuals of respondents within a cluster would distribute closely in a multidimensional space. A preliminary tool for assessing the homogeneity of residuals is required. Cluster analysis, which is a set of multivariate techniques for grouping objects with similar characteristics into clusters, may be helpful for dealing with such a task. Cluster analysis can be divided roughly into two branches: hierarchical and nonhierarchical. Hierarchical cluster analysis (HCA) provides a dendrogram displaying that an observation or cluster of observations is nested under another cluster, whereas nonhierarchical cluster analysis (i.e., k-means) simply divides observations into several clusters. A combination of using the

hierarchical approach to determine the number of clusters, followed by the nonhierarchical approach to achieve more accurate cluster memberships, is often advisable in practice.

Two issues deserve notice when assessing the magnitude of LPD by means of cluster analysis. First, the number of clusters and the whole sample size should be jointly considered in cluster analysis. It is inefficient to include whole observations in the analysis. The number of explored clusters increases, along with the number of observations, but is not necessarily identical to the number of true clusters. Subsequently, the correspondence between the explored clusters and true clusters would be all in a mass. It seems applicable to include a set of observations within two or three true clusters in an analysis. Another concern is related to screening the analyzed sample. To highlight the homogeneity of residuals for respondents within a cluster and the heterogeneity between clusters, matching clusters according to cluster-level scores (i.e., GPD) is strongly recommended. Suppose two clusters are selected: One has extremely high scores and the other one has extremely low scores. The results from cluster analysis would show that explored clusters are akin to true clusters, even though person residuals are mutually independent of each other, because the influence of GPD is not excluded when investigating the existence of LPD.

HCA, which provides a visual and intuitive output, was selected in this study for detecting LPD. The step-by-step procedure of HCA for detecting LPD was illustrated as follows:

1. Compute the mean raw score for each cluster;
2. Match clusters with similar mean scores;
3. Choose moderate size of samples from the matched clusters;
4. Fit a standard model to data;
5. Compute the residuals of selected samples as the analyzed variables in HCA;
6. Observe the coincidence between the explored clusters and true clusters.

When the selected samples can be distinctively grouped in the two explored clusters, the existence of LPD can be confirmed.

Chapter 4: Simulation Studies

Several simulation studies are presented in this chapter. The following simulations focused on polytomous models because dichotomous models are included as special cases. The findings under polytomous models can be generalized to those under dichotomous models. To understand the applicability of the new models to unidimensional test items, the parameter recovery for the MPCM-CS is evaluated, and the consequences of ignoring LPD are of interest. Subsequently, the influences of cluster numbers and within-cluster sample sizes on parameter estimation are investigated. The parameter recovery for the two generalized models (i.e., the MDMPCM-CS and HOPCM-CS) for tests composed of multiple scales is also examined. Because multi-faceted data (e.g., ratee, rater, and item) is not rare, it is necessary to test whether the parameters in the MFRM-CS could be recovered accurately. In addition to modeling LPD directly, simulated responses are used to investigate the level of efficiency for HCA. These issues are addressed in sequence.

4.1 Simulation study 1: Parameter recovery of MPCM-CS

Study 1 focused on the parameter recovery for the MPCM-CS and on the consequences of ignoring LPD in parameter estimation. The settings were designed according to the scenario in a large-scale assessment. There were 10 polytomous items. The mean-item difficulties were randomly generated from U (-1.5, 1.5), and the step parameters were set at -0.5, 0, and 0.5 for each item, i.e., the item thresholds were between -2 and 2. In the LPD condition, item responses were generated from the MPCM-CS, and the 10 values of $\sigma^2_{\xi_i}$ were set at 0.4, 0.8, and 1.2 for four, three, and three items, respectively. Conversely, in the nil condition, item responses were generated from the MPCM (a simpler model without ξ parameters) so that the 30 values of $\sigma^2_{\xi_i}$, by definition, were all zero. Two sample sizes were considered. In the small-sample condition, 1,000 persons were sampled from 100 clusters (e.g., schools) with 10 persons in each cluster, whereas in the large-sample condition, 2,000

persons were sampled from 200 clusters (e.g., schools) with 10 persons in each cluster. Latent variable θ was generated via a random-intercept model with one binary predictor (e.g., male vs. female) at Level 1 and one binary predictor (e.g., public school vs. private school) at Level 2. The mean level of the measured latent trait was -0.3 and 0.3 for males and females, respectively, and -0.2 and 0.2 for public and private schools, respectively. The explained variance by clusters was set at 0.25, and the residual variance was set at 0.64, i.e., the ICC was set at .281. Both the MPCM-CS and MPCM were fit to the generated data.

Each condition included a total of 100 replications. The following priors were adopted in WinBUGS: $N(0, 10)$ for the item difficulties, the regression coefficients of Level 1 and Level 2 predictors, and $\Gamma(0.1, 0.1)$ for the inverse of variances, including σ_{υ}^2, σ_{ε}^2, and $\sigma_{\xi_i}^2$. The WinBUGS codes for the MPCM-CS can be found in Appendix A. The bias and root mean square error (RMSE) were computed for each parameter:

$$\text{Bias}\left(\hat{\xi}\right) = \sum_{r=1}^{R}\left(\hat{\xi} - \zeta\right) \Big/ R, \tag{37}$$

$$\text{RMSE}\left(\hat{\xi}\right) = \sqrt{\sum_{r=1}^{R}\left(\hat{\xi} - \zeta\right)^2 \Big/ R}, \tag{38}$$

in which ζ and $\hat{\xi}$ are the true value and parameter estimate, respectively, and R denotes the number of replications (i.e., 100). Three research hypotheses were examined.

1. In the LPD conditions, the parameters in the MPCM-CS can be recovered very well, while ignoring LPD by fitting the MPCM results in biased parameters; in the nil conditions, fitting the new model to data without LPD would still yield good parameter recovery.
2. In comparing the first two conditions, it is clear that a larger sample size can improve the performance of the MPCM-CS, especially for ξ parameters.
3. Given the identical total sample size (i.e., conditions 3 and 4), more clusters lead to better

recovery for parameters regarding the multilevel structure.

Table 1 summarizes the bias and RMSE values when the MPCM and MPCM-CS were applied to the data simulated from the MPCM-CS. As expected, the MPCM yielded poor estimation, whereas the MPCM-CS recovered the parameters very well. When the MPCM was fit, the bias for the threshold parameters was between -0.775 and 0.612, and the RMSE was between 0.105 and 0.787 when the sample size was 1,000. Meanwhile, the bias was between -0.902 and 0.762, and the RMSE was between 0.087 and 0.907 when the sample size was 2,000. The results showed that using a large sample size does not improve parameter recovery because LPD was not considered. Figures 1a and 1b illustrate the patterns of biased estimation under the MPCM. When item responses were contaminated by LPD, item thresholds with positive values were generally underestimated, whereas those with negative values were generally overestimated, suggesting that LPD resulted in a shrunken scale in the MPCM.

The shrunken scale in the MPCM would, in turn, influence the parameter estimates of Level 1 and Level 2 variances: The bias for σ_{ε}^2 and σ_u^2 was -0.387 and -0.097, respectively, when the sample size was 1,000, and it was -0.385 and -0.101 when the sample size was 2,000. The aforementioned shrinkage of the scale was due to the fact that LPD was not considered in the MPCM. Due to the underestimation of σ_{ε}^2 and σ_u^2, the group differences in Level 1 and Level 2 shrank toward the prior mean of zero. Furthermore, the shrunken scale caused by LPD would influence the estimation of the ICC. For example, the MPCM yielded ICCs of .377 and .369 when the sample size was 1,000 and 2,000, respectively; therefore, ignoring LPD inflates GPD.

Table 1. Summary of parameter recovery for the MPCM and MPCM-CS under the LPD condition in simulation study 1.

	MPCM				MPCM-CS			
	$N = 1000$		$N = 2000$		$N = 1000$		$N = 2000$	
	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
Threshold (δ)								
Max	0.612	0.787	0.762	0.907	0.027	0.202	0.018	0.152
Min	-0.775	0.105	-0.902	0.087	-0.024	0.111	-0.013	0.086
Mean	-0.102	0.396	-0.010	0.403	-0.003	0.147	0.003	0.107
LPD (σ^2_ξ)								
Max	–	–	–	–	0.060	0.246	0.035	0.184
Min	–	–	–	–	0.010	0.099	-0.012	0.072
Mean	–	–	–	–	0.034	0.167	0.015	0.118
Level 1								
η	-0.100	0.102	-0.099	0.100	0.001	0.031	0.001	0.021
σ^2_ε	-0.387	0.387	-0.385	0.385	0.010	0.045	0.009	0.034
Level 2								
γ	-0.057	0.074	-0.066	0.074	0.014	0.068	0.002	0.051
σ^2_u	-0.097	0.101	-0.101	0.103	0.010	0.058	0.002	0.042

Note. N denotes the sample size; – = not applicable.

Table 2. Summary of parameter recovery for the MPCM and MPCM-CS under the nil condition in simulation study 1.

	MPCM				MPCM-CS			
	N = 1000		*N* = 2000		*N* = 1000		*N* = 2000	
	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
Threshold (δ)								
Max	0.020	0.181	0.016	0.128	0.109	0.208	0.077	0.147
Min	-0.029	0.094	-0.018	0.066	-0.071	0.094	-0.070	0.066
Mean	-0.001	0.124	-0.003	0.088	0.019	0.135	-0.002	0.095
LPD (σ^2_ξ)								
Max	–	–	–	–	0.061	0.062	0.046	0.047
Min	–	–	–	–	0.048	0.049	0.035	0.035
Mean	–	–	–	–	0.054	0.055	0.040	0.041
Level 1								
η	-0.001	0.030	0.002	0.020	0.009	0.032	0.010	0.022
σ^2_ε	0.008	0.046	-0.003	0.030	0.058	0.076	0.035	0.047
Level 2								
γ	-0.007	0.059	-0.005	0.042	0.000	0.060	0.000	0.042
σ^2_u	0.007	0.051	0.002	0.031	0.056	0.057	0.011	0.034

Note. *N* denotes the sample size; – = not applicable.

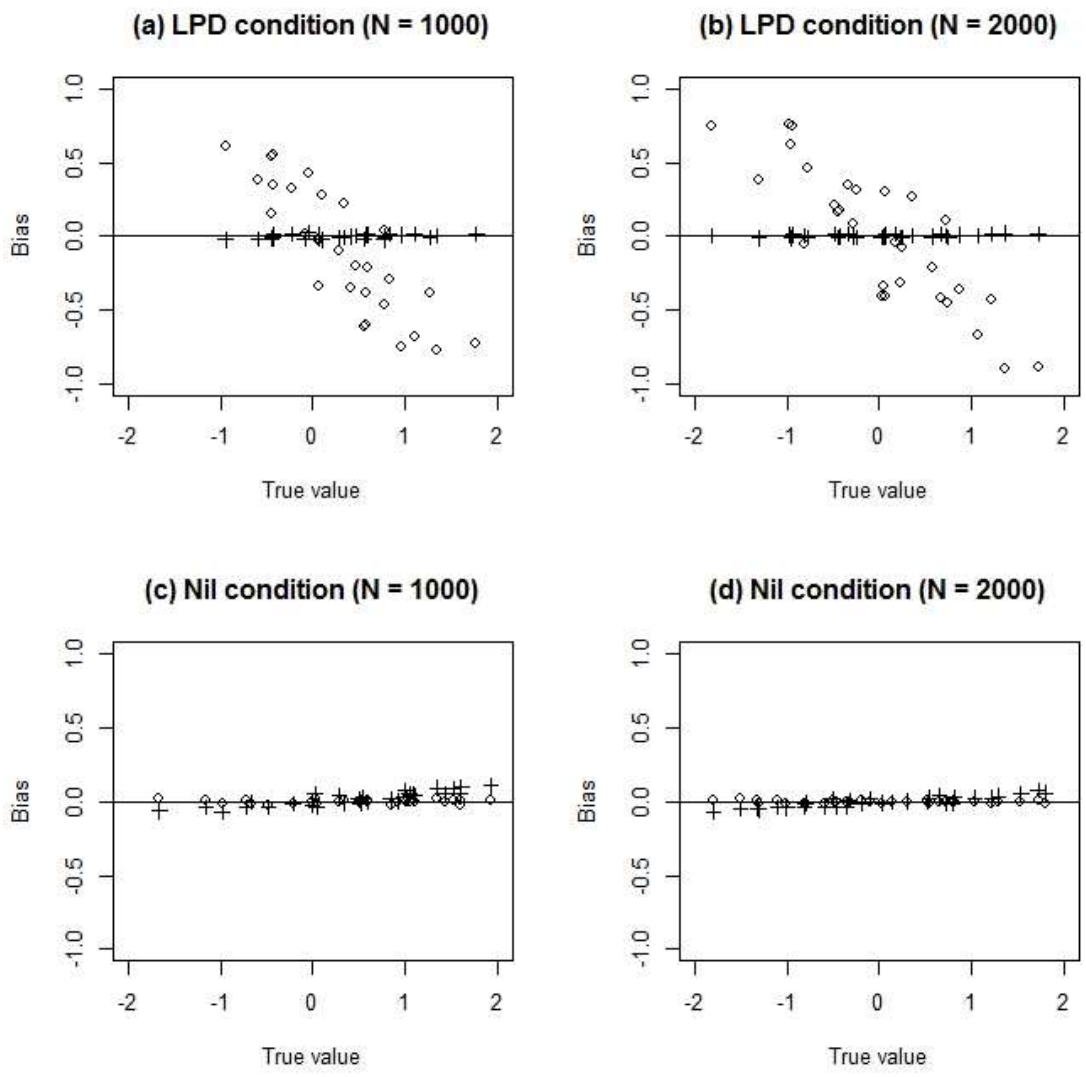

Figure 1. Bias for the threshold parameters under the MPCM and MPCM-CS

Note. Circles denote the biases under the MPCM, and crosses are the biases under the MPCM-CS.

On the other hand, all the estimated parameters were recovered accurately in the MPCM-CS: The bias was between -0.024 and 0.027, and the RMSE was between 0.111 and 0.202 when the sample size was 1,000. In addition, the bias was between -0.013 and 0.018, and the RMSE was between 0.086 and 0.152 when the sample size was 2,000. Regarding the estimate of σ_{ξ}^{2}, the bias was between 0.010 and 0.060, and the RMSE was between 0.099

and 0.246 when the sample size was 1,000. Furthermore, the bias was between -0.012 and 0.035, and the RMSE was between 0.072 and 0.184 when the sample size was 2,000. This suggested that the estimate of σ_{ξ}^2 would be very close to zero with a larger sample size. Hence, the MPCM-CS was able to detach the nuisance of LPD and recover the parameters very well.

In terms of test reliability, the mean estimate in the MPCM was.785 and .784 when the sample size was 1,000 and 2,000, respectively. Meanwhile, the mean estimate in the MPCM-CS was .811 when the sample size was 1,000 and 2,000. It appeared that ignoring LPD tended to lead to the underestimating of test reliability slightly.

Table 2 summarizes the bias and RMSE values when the MPCM and MPCM-CS were fit to the simulated data without LPD. Both models yielded unbiased estimates for the item difficulties. As shown in Figures 5c and 5d, the item difficulties were recovered fairly well by fitting the MPCM or MPCM-CS. Although σ_{ε}^2 and σ_{u}^2 were upwardly estimated when the MPCM-CS was fit, the magnitudes of the overestimation were trivial and acceptable.

In sum, fitting the unnecessarily complicated MPCM-CS to data without LPD did little harm to the parameter estimation and yielded close-to-zero estimates for σ_{ξ}^2. Ignoring LPD by fitting the MPCM-CS yielded a shrunken scale, biased estimates for the item parameters, and deflated test reliability.

4.2 Simulation study 2: Influence of different combinations of numbers of clusters and within-cluster sample sizes

Following the designs in simulation study 1, the influence of different combinations of clusters and intra-cluster sample sizes -- conditional on a fixed total sample size -- on parameter recovery was examined in simulation study 2. Two sample structures were generated. In the few-cluster condition, 1,000 persons were sampled from 20 clusters featuring 50 persons each, whereas in the multi-cluster condition, 1,000 persons were

sampled from 50 clusters featuring 20 persons each. Both scenarios are realistic. Item responses were generated from the MPCM-CS, and the parameter settings were the same as those in simulation study 1. The true model (i.e., MPCM-CS) was fit to the generated data. In multilevel modeling, the effective sample size (ESS; Kish, 1965), which influences the precision of estimation, is particularly of concern:

$$\mathrm{ESS} = \frac{N_{total}}{1 + \mathrm{ICC} \times \left(N_{cluster} - 1\right)}, \tag{39}$$

in which N_{total} is the actual sample size, and $N_{cluster}$ is the number of clusters. According to Kish's formula, given a fixed total sample size and a positive ICC, a larger ESS would be achieved for a dataset composed of more clusters. Three research hypotheses were examined.

1. Using more clusters is helpful for recovering the LPD parameters, as well as Level 2 parameters, because a larger number of clusters contribute more information for estimating these parameters.
2. The precision of the recovery for item parameters might be lessened by the poorer estimation of the LPD parameters when there are fewer clusters.
3. The recovery for Level 1 parameters under the two conditions would be very similar due to the constant sample size.

Table 3 summarizes the bias and RMSE values for 20 clusters and 50 clusters. As expected, the recovery for the LPD parameters was better under the condition of 50 clusters. For the LPD parameters, the bias was between 0.067 and 0.176, and the RMSE was between 0.196 and 0.571 when there were 20 clusters. Meanwhile, the bias was between -0.008 and 0.087, and the RMSE was between 0.116 and 0.338 when there were 50 clusters. Poorer recovery was found for Level 2 parameters when there were 20 clusters. For example, the RMSE for γ_{01} (e.g., the difference between public and private schools) was 0.120 when there were 20 clusters, and it was 0.070 when there were 50 clusters. In line with expectations, the

magnitudes of the biased estimates for item parameters were salient under the condition of 20 clusters, compared with the results from the condition of 50 clusters. The results also indicated that the number of clusters did not influence the estimation of Level 1 parameters. Overall, the results matched the expectation that the larger the number of clusters, the better the parameter estimation would be, given a fixed total sample size.

Table 3. Summary of parameter recovery under different sampling structure conditions in simulation study 2.

	20 clusters		50 clusters	
	Bias	RMSE	Bias	RMSE
Threshold (δ)				
Max	0.051	0.326	0.038	0.247
Min	-0.072	0.180	-0.049	0.124
Mean	-0.012	0.255	-0.002	0.185
LPD (σ^2_ξ)				
Max	0.176	0.571	0.087	0.338
Min	0.067	0.196	-0.008	0.116
Mean	0.122	0.365	0.050	0.217
Level 1				
η	0.007	0.030	0.001	0.027
σ^2_ε	0.005	0.048	0.005	0.048
Level 2				
γ	-0.025	0.132	-0.001	0.075
σ^2_u	0.048	0.120	0.011	0.070

4.3 Simulation study 3: Parameter recovery of the MDMPCM-CS

Study 3 was aimed at examining the parameter recovery for the MDMPCM-CS and the consequences of fitting the MDMPCM (a simpler model without ξ parameters). There were three tests measuring different, but correlated, latent traits in each. Five four-point items were included in each test. Like the settings in simulation study 1, the item thresholds were between -2 and 2. In each test, the five values of $\sigma^2_{\xi_i}$ were set at 0.2, 0.4, 0.6, 0.8, and 1,

respectively. A sample of 2,000 persons was generated, and all examinees were divided into 200 clusters, with each cluster containing 10 persons. Due to the lengthy computation time (approximately 10 hours per replication), a larger sample was not considered in this simulation. The covariance matrix for the individual-level effects was set as $\begin{bmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{bmatrix}$, and for the cluster-level effects, it was set as $\begin{bmatrix} 1 & 0.2 & 0.2 \\ 0.2 & 1 & 0.2 \\ 0.2 & 0.2 & 1 \end{bmatrix}$. In addition, there was one binary predictor (e.g., male vs. female) at Level 1, and the group differences on these three dimensions were set at 0.2, 0.4, and 0.6. Both the MDMPCM-CS and MDMPCM were fit to the generated data. The WinBUGS codes for the MDMPCM-CS can be found in Appendix B. The simulation carried out a total of 100 replications. After fitting the data-generating model, the bias and RMSE in the parameter estimates were computed to evaluate parameter recovery. It was expected that the parameters in the MDMPCM-CS can be recovered very well, but fitting the MDMPCM leads to biased estimation.

Table 4 summarizes the bias and RMSE values when the MDMPCM and MDMPCM-CS were applied to the simulated MDMPCM data. Similar to the findings in study 1, the MDMPCM-CS could recover all the estimated parameters very well. For the item thresholds, the bias was between -0.037 and 0.039, and the RMSE was between 0.100 and 0.157. For the LPD parameters, the bias was between -0.019 and 0.043, and the RMSE was between 0.057 and 0.183. For the Level 1 group difference, the bias was between -0.005 and 0.000, and the RMSE was roughly 0.028. For the covariance matrices in Level 1 and Level 2, the bias was between -0.010 and 0.019, and the RMSE was between 0.038 and 0.139. In general, the parameter recovery was satisfactory.

When the MDMPCM was fit to the data generated from the MDMPCM-CS, the

estimates of the item thresholds were far away from their true values and shrank toward the mean. The bias was between -0.914 and 0.790, and the RMSE was between 0.089 and 0.922. The parameters of the two covariance matrices were consistently underestimated: The bias was between -0.539 and -0.092, and the RMSE was between 0.104 and 0.540. Furthermore, after standardizing the covariance matrices, it was found that the correlation between pairs of random variables in Level 1 had a mean inflation of 0.08, whereas the correlation between pairs of random variables in Level 2 had a mean deflation of 0.03. The group differences in the random variables in Level 1 were downwardly estimated.

Table 4. Summary of parameter recovery for the MDMPCM and the MDMPCM-CS in simulation study 2.

	MDMPCM		MDMPCM-CS	
	Bias	RMSE	Bias	RMSE
Threshold (δ)				
Max	0.790	0.922	0.020	0.157
Min	-0.914	0.089	-0.036	0.100
Mean	-0.003	0.363	-0.004	0.126
LPD (σ^2_ξ)				
Max	–	–	0.043	0.183
Min	–	–	-0.019	0.057
Mean	–	–	0.015	0.114
Level 1				
$\boldsymbol{\eta}$				
Max	-0.031	0.086	0.000	0.028
Min	-0.084	0.036	-0.005	0.028
Mean	-0.056	0.060	-0.003	0.028
$\boldsymbol{\Sigma}_\varepsilon$				
Max	-0.226	0.540	0.018	0.071
Min	-0.539	0.227	0.003	0.038
Mean	-0.382	0.383	0.010	0.054
Level 2				
$\boldsymbol{\Sigma}_\mathbf{u}$				
Max	-0.092	0.401	0.019	0.139
Min	-0.394	0.104	-0.010	0.092
Mean	-0.243	0.253	0.002	0.115

4.4 Simulation study 4: Parameter recovery of the HOPCM-CS

Study 4 was aimed at examining the parameter recovery for the HOPCM-CS and the consequence of fitting the HOPCM (a simpler model without ξ parameters). There were four first-order latent traits and one second-order latent trait. Each test had five four-point items. As with the settings in simulation study 1, the item thresholds were between -2 and 2. In each test, the five values of $\sigma^2_{\xi_i}$ were set at 0.2, 0.4, 0.6, 0.8, and 1. The factor loadings, which

represent the relationships between the second-order latent trait and the four first-order latent traits, were set at 1, 0.9, 0.8, and 0.7, and all the variances of the regressed residuals were set at 0.16 for the four first-order latent traits, i.e., the variances explained by the second-latent trait on the four first-order latent traits were 0.86, 0.84, 0.80, and 0.75. A sample of 2,000 persons was generated, and all examinees were divided into 200 clusters, with each school having 10 persons. Likewise, there was one binary predictor (e.g., male vs. female) at Level 1 and one binary predictor (e.g., public school vs. private school) at Level 2. The mean level of the second-order latent trait was -0.3 and 0.3 for males and females, respectively, and the public and private schools had different mean levels for the latent trait, at -0.2 and 0.2, respectively. The explained variance by clusters on the second-order latent trait was set at 0.25, and the residual variance was set at 0.64. The WinBUGS codes for the HOPCM-CS can be found in Appendix C. The simulation included a total of 100 replications. After fitting the data-generating model, the bias and RMSE in the parameter estimates were computed to evaluate parameter recovery. It was expected that the parameters in the HOPCM-CS can be recovered very well, but fitting the HOPCM leads to biased estimations.

Table 5 summarizes the bias and RMSE values when the HOPCM and HOPCM-CS were applied to the simulated HOPCM data. Similar to the findings in study 1, the HOPCM-CS could recover all the estimated parameters very well. For the item thresholds, the bias was between -0.036 and 0.020, and the RMSE was between 0.073 and 0.153. For the factor loadings, the bias was -0.005, and the RMSE was between 0.033 and 0.037. For the LPD parameters, the bias was between -0.003 and 0.034, and the RMSE was between 0.039 and 0.168. For the other parameters, the bias was between 0.000 and 0.013, and the RMSE was between 0.021 and 0.050. In general, the parameter recovery was satisfactory. When the HOPCM was fit to the data generated with LPD, the estimates of the item thresholds were far from their true values and shrank toward the prior mean. The bias was between -0.891 and

1.020, and the RMSE was between 0.071 and 1.029. In addition, the relationship between the first- and second-order latent traits was distorted. Although the loadings were precisely recovered, all the residuals were downwardly biased. After standardizing the loadings, the empirical relationships were overestimated. Finally, the HOPCM yielded a shrunken scale of the second-order latent trait. The variances in Level 1 and Level 2 were substantially underestimated, and the group differences in Level 1 and Level 2 were attenuated.

Table 5. Summary of parameter recovery for the HOPCM and the HOPCM-CS in simulation study 4.

	HOPCM		HOPCM-CS	
	Bias	RMSE	Bias	RMSE
Threshold (δ)				
Max	1.020	1.029	0.020	0.153
Min	-0.891	0.071	-0.036	0.073
Mean	0.043	0.367	-0.004	0.104
Loadings (λ)				
Max	0.012	0.052	-0.005	0.037
Min	-0.013	0.041	-0.005	0.033
Mean	-0.003	0.045	-0.005	0.035
Residuals (ν)				
Max	-0.071	0.079	0.006	0.029
Min	-0.077	0.073	0.000	0.021
Mean	-0.074	0.076	0.003	0.026
LPD (σ^2_ξ)				
Max	–	–	0.034	0.168
Min	–	–	-0.003	0.039
Mean	–	–	0.011	0.098
Level 1				
η	-0.083	0.084	0.002	0.022
σ^2_ε	-0.325	0.326	0.013	0.050
Level 2				
γ	-0.054	0.063	0.003	0.046
σ^2_u	-0.091	0.094	0.012	0.045

4.5 Simulation study 5: Parameter recovery of the MFRM-CS

Study 5 tested the parameter recovery for the MFRM-CS. Because three-faceted data (including the ratee, item, and rater) are very common, in this study, the LPD, as the ratee-rater interaction, was of concern. Item responses were generated from the MFRM-CS. There were five criteria (items) rated in five response categories. Two levels of saturation of the data matrix were manipulated. There were 10 raters, and each ratee was marked by either

two or four of the 10 raters, suggesting 80% and 60% missingness, respectively. The overall difficulties were set at -1, -0.5, 0, 0.5, and 1 for the five criteria, and the four threshold difficulties were set at -0.75, -0.25, 0.25, and 0.75 for all criteria. The rater severities were sampled from the normal distribution, with a mean of zero and a variance of 0.36, and the first rater was constrained at zero for identification. The 10 values of $\sigma^2_{\xi_k}$ were set at 0.4, 0.8, and 1.2 for four, three, and three raters, respectively. The simulation included 200 clusters, and each cluster included 10 persons. The settings of the latent variable θ were identical to those in simulation study 1. The WinBUGS codes for the MFRM-CS can be found in Appendix D. Likewise, there were 100 replications, and the bias and RMSE in the parameter estimates were computed for evaluating parameter recovery. Although the generated data were fit via the data-generating model, the parameter estimates may not be satisfactory because the data matrix may be too sparse to provide sufficient information for estimation. It was expected that, with less-sparse data (i.e., a ratee was rated by more raters), the estimates would be closer to their true values.

Table 6 summarizes the bias and RMSE values when the MFRM-CS was applied to the multifaceted MFRM-CS data. When each ratee was rated by two raters, for the item thresholds, the bias was between -0.001 and 0.024, and the RMSE was between 0.132 and 0.167. For rater severities, the bias was between -0.024 and 0.020, and the RMSE was between 0.180 and 0.278. For the LPD parameters, the bias was between -0.030 and 0.086, and the RMSE was between 0.137 and 0.399. Finally, for the other parameters, the bias was between -0.006 and 0.053, and the RMSE was between 0.002 and 0.107. When each ratee was rated by four raters, for the item thresholds, the bias was between 0.002 and 0.017, and the RMSE was between 0.094 and 0.120. For the rater severities, the bias was between -0.015 and 0.015, and the RMSE was between 0.112 and 0.155. For the LPD parameters, the bias was between 0.013 and 0.064, and the RMSE was between 0.085 and 0.250. For the other

parameters, the bias was between -0.002 and 0.011, and the RMSE was between 0.020 and 0.058. These results support the expectation that the parameters in the MFRM-CS could be accurately recovered, and the recovery was even better with more saturated data.

Table 6. Summary of parameter recovery for the MFRM-CS in simulation study 5.

	Two raters		Four raters	
	Bias	RMSE	Bias	RMSE
Threshold (δ)				
Max	0.024	0.167	0.017	0.120
Min	-0.001	0.132	0.002	0.094
Mean	0.012	0.145	0.010	0.105
Severity (ι)				
Max	0.020	0.278	0.015	0.155
Min	-0.024	0.180	-0.015	0.112
Mean	-0.004	0.237	0.003	0.136
LPD (σ^2_ξ)				
Max	0.086	0.399	0.064	0.250
Min	-0.030	0.137	0.013	0.085
Mean	0.029	0.246	0.030	0.161
Level 1				
η	-0.001	0.002	0.003	0.020
σ^2_ε	0.007	0.032	-0.002	0.030
Level 2				
γ	-0.006	0.107	0.005	0.058
σ^2_u	0.053	0.099	0.011	0.055

4.6 Performance of HCA

A total of 1,000 persons (sampled from 100 clusters featuring 10 persons each), who responded to 10 four-point items, were generated from the MPCM-CS. Both the MPCM and MPCM-CS were fit to the generated data. The reader is referred to Chapter 3.5 for parameter estimation and Chapter 3.6 for more details on data generation. The total score of the generated clusters was between 47 and 225. HCA was conducted by using R 3.2.5 (R Core Team, 2016). The dissimilarities between clusters were the squared Euclidean distances

between cluster means, and Ward's (1963) clustering criteria were implemented to choose the pair of clusters to merge at each step. It was expected that, when the residuals from the MPCM were analyzed, a high correspondence between the explored and true clusters would be found. In contrast, when the residuals from the MPCM-CS were analyzed, observations would be randomly distributed in the explored clusters.

Four combinations were presented: (a) Clusters 4 and 8 had a mean score of 8.3; (b) clusters 38 and 57 had a mean score of 10.9; (c) clusters 1 and 86 had a mean score of 12.9; and (d) clusters 51, 53, and 78 had a mean score of 10.3. The first three combinations included two clusters, and the last combination included three clusters.

Figures 2a and 3a show good recovery of the true clusters when the residuals from the MPCM were analyzed. This suggested that the standard model did not fully account for LPD, so the respondents within a cluster were accurately grouped based on the similarity of the residuals. On the other hand, when the residuals from the MPCM-CS were analyzed, systematic patterns of grouping were not found (Figures 2b and 3b). Figure 4 shows the results when HCA was applied to three clusters. It seems that the generated simulees within the same clusters could be grouped roughly according to the residuals from the MPCM. Figure 5, however, displays a flaw, in which HCA could not distinguish the observations clearly. The extent of heterogeneity of residuals between clusters might be an explanation for why HCA sometimes performed well, but other times did not. When selected clusters exhibit similar LPD patterns, limited information can be provided by the person residuals from the standard model to identify true clusters, leading to poor HCA performance. Computing the Euclidean distance between two vectors of LPD estimates for two clusters is helpful for comprehension. The distance for the two good-classification pairs (Figures 2 and 3) was 4.05 and 4.22, and for the poor-classification pair (Figure 5), the distance was 2.51. In other words, although the magnitudes of LPD were substantial in general, they might not have been

detected if only a subset of clusters were selected. In sum, HCA did not perform consistently in detecting LPD. To achieve better detection of LPD, the proposed models are strongly recommended.

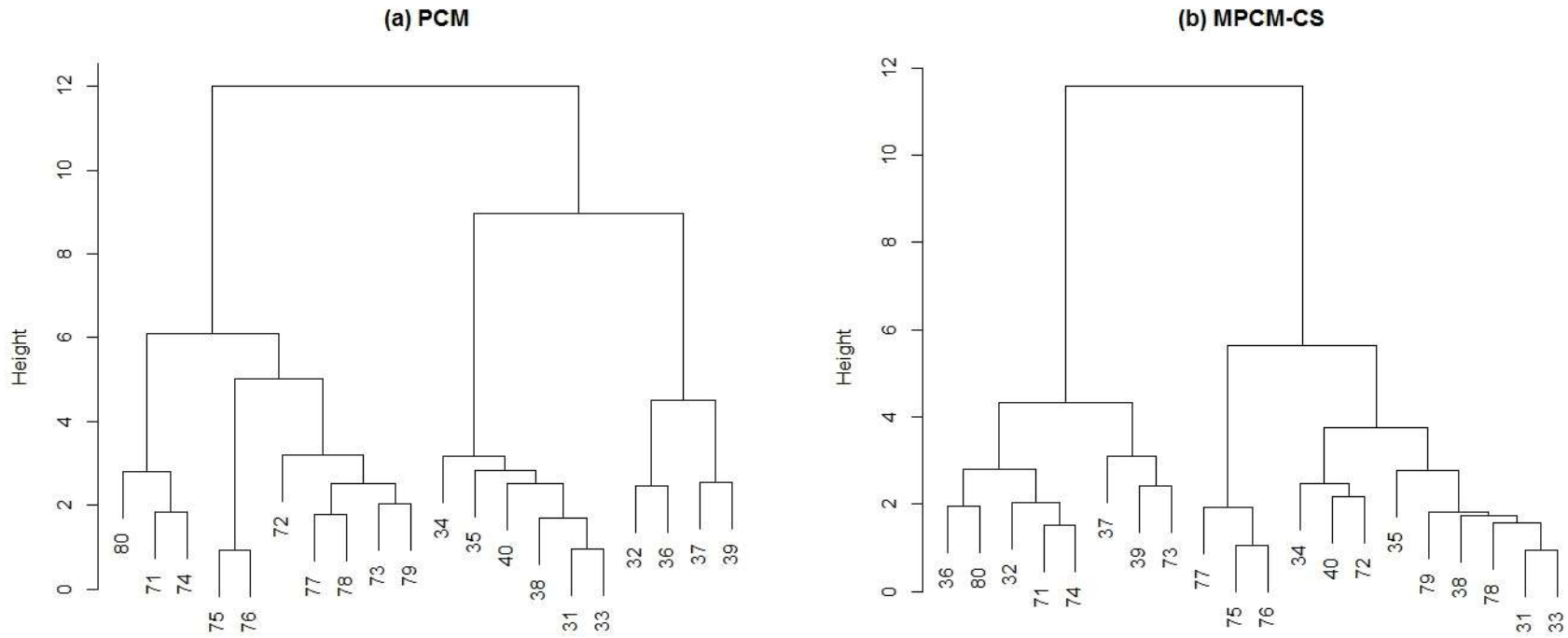

Figure 2. Dendrograms for the first example of good classification.

Note. Cluster 4 includes simulees 31-40; and cluster 8 includes simulees 71-80.

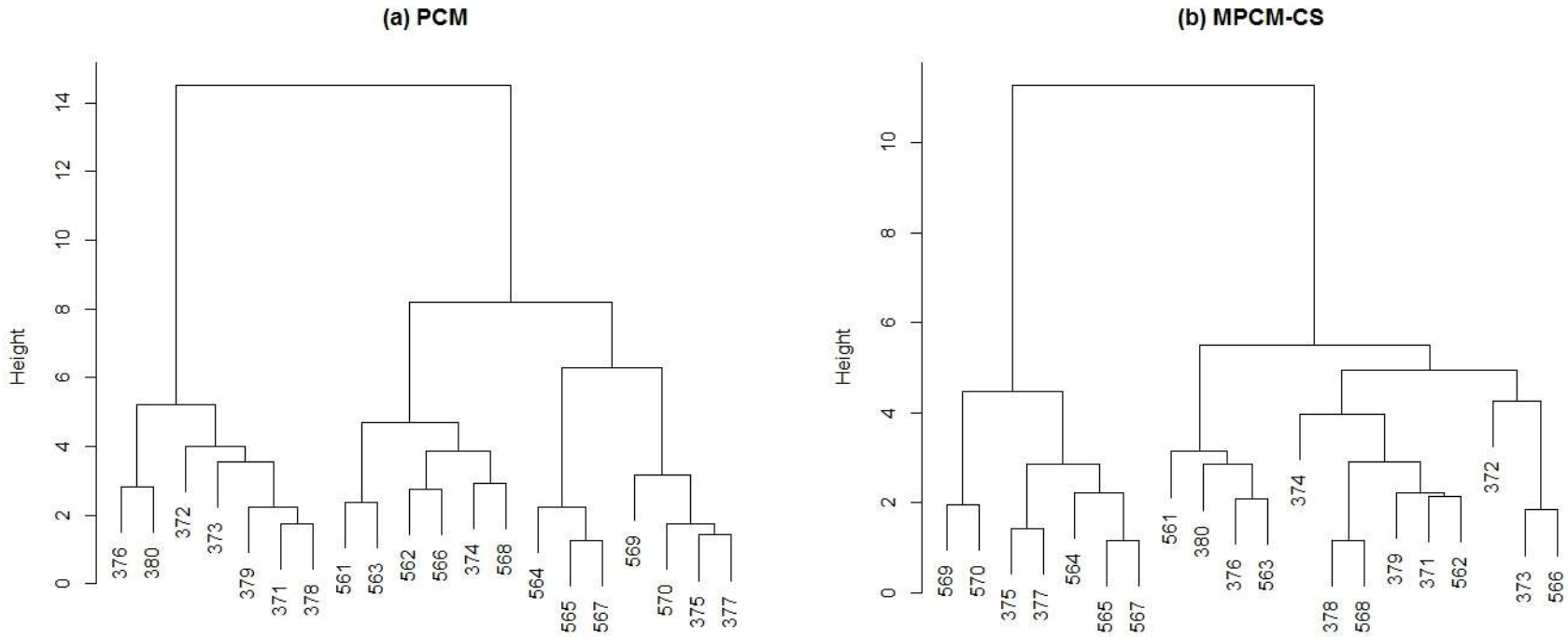

Figure 3. Dendrograms for the second example of good classification.

Note. Cluster 38 includes simulees 371-380; and cluster 57 includes simulees 561-570.

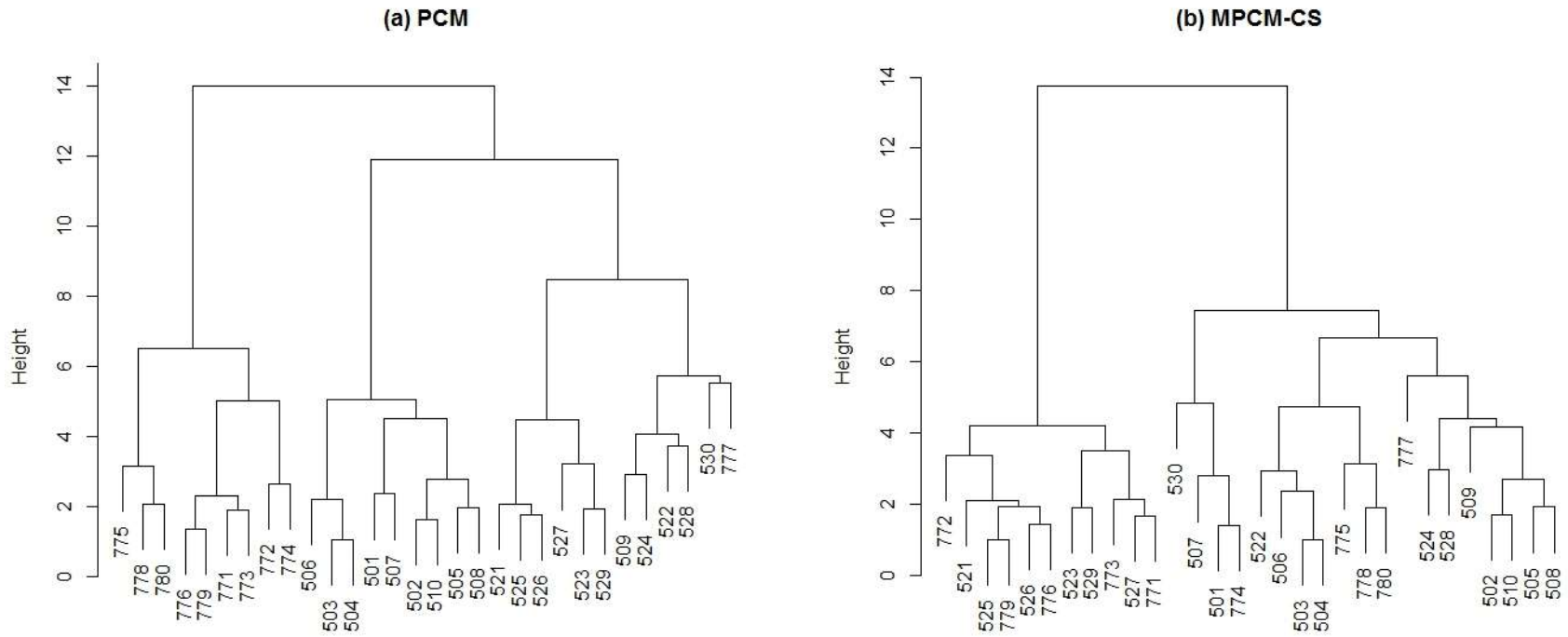

Figure 4. Dendrograms for the example of good classification for three clusters.

Note. Cluster 51 includes simulees 501-510; cluster 53 includes simulees 521-530; and cluster 78 includes simulees 771-780.

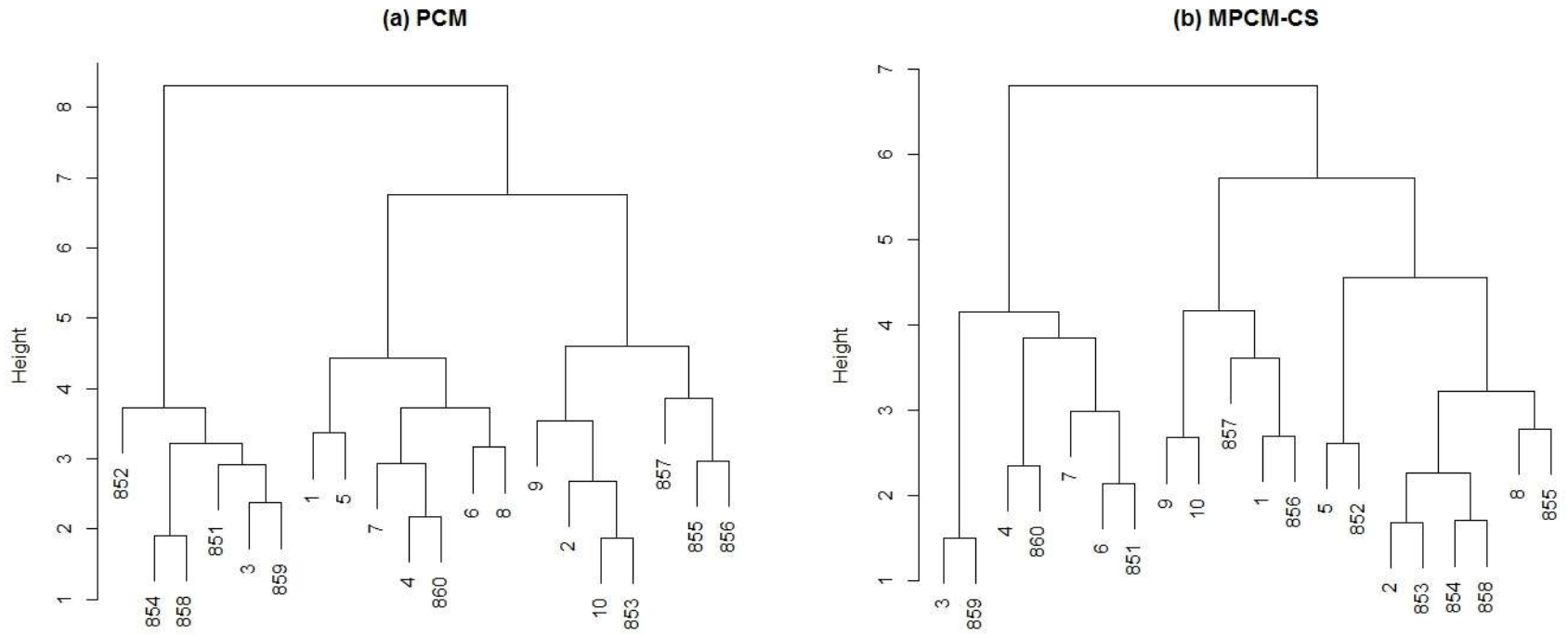

Figure 5. Dendrograms for the example of poor classification.

Note. Cluster 1 includes simulees 1-10; and cluster 86 includes simulees 851-860.

Chapter 5: Empirical Example Studies

5.1 Example 1: General knowledge in daily life

The first example is the National Longitudinal Study of Adolescent Health (Add Health) (Harris & Udry, 2016) project, which aims to investigate how social environments and behaviors in adolescence influence health and achievement outcomes in young adulthood. A subscale in the project, surveying whether adolescents have learned the listed items with respect to general knowledge in daily life at school, was adopted. Recruited adolescents were in the seventh to 11th grades and were stratified and randomly sampled from all high schools in the U.S. The analysis included 6,493 cases from 143 schools. Each school included 12 to 122 adolescents. Both the MRM and MRM-CS were fit to the complete reading data. Because schools in the U.S. can choose different versions of textbooks, and because the regional differences in the influence of urbanization and education investment are considerable, the taught contexts across schools might be very divergent. Accordingly, it is intuitive to expect that adolescents within the same school are more homogeneous than students from different schools, both in overall levels of learning general knowledge (i.e., GPD) and specific subject matter (i.e., LPD) emphasized by their schools. Fitting the MRM-CS is effective for quantifying different kinds of dependences. Furthermore, HCA was conducted to investigate the homogeneity of the person residuals obtained from the MRM and MRM-CS. Schools were paired based on mean scores of adolescents on 17 items, and two schools at similar levels were selected. Fifteen adolescents were randomly sampled from each school for the subsequent HCA. One could anticipate that adolescents within a school would be grouped together if the magnitudes of LPD were substantial.

Table 7 summarizes the mean and standard deviation of the raw scores and the mean thresholds under the MRM and MRM-CS for each item. It shows that most content with respect to general knowledge had been broadly taught at schools. The item chosen least

often was “The problems of being underweight,” whereas the item chosen the most was “Drug abuse.” This also shows that the MRM yielded a narrower range than did the MRM-CS. Considering the findings in simulation study 1, the current results indicated substantial LPD.

The AIC, BIC, and DIC for the MRM were 75,810, 75,939, and 80,601, respectively, and for the MRM-CS, they were 72,327, 72,571, and 78,262, respectively, indicating that the MRM-CS yielded a better model-data fit than did the MRM. The posterior predictive *p*-value for the MRM and the MRM-CS was 0.990 and 0.972, respectively. Because there were 4,393 adolescents and 17 items, if all person residuals were independent, the expected value and the standard deviation of Q_3 would be -0.0001 and 0.258, respectively. The empirical mean Q_3 was 0.032 for the MRM and 0.026 for the MRM-CS; the empirical standard deviation was 0.366 for the MRM and 0.344 for the MRM-CS, suggesting that the MRM-CS had a better fit, according to the Q_3 statistic.

Table 7. Raw scores and difficulty estimates under the MRM and MRM-CS in the general knowledge scale.

		Raw score		MRM		MRM-CS	
No.	Item	Mean	SD	Estimate	SE	Estimate	SE
1	The foods you should and shouldn't eat	0.870	0.337	-2.587	0.075	-2.754	0.094
2	The importance of exercise	0.922	0.269	-3.302	0.084	-3.481	0.108
3	Smoking	0.919	0.273	-3.265	0.081	-3.461	0.104
4	The problems of being overweight	0.596	0.491	-0.522	0.069	-0.536	0.087
5	Drinking	0.938	0.240	-3.625	0.082	-3.870	0.118
6	Drug abuse	0.956	0.205	-4.045	0.090	-4.296	0.121
7	Pregnancy	0.860	0.347	-2.480	0.075	-2.571	0.106
8	AIDS	0.919	0.273	-3.260	0.083	-3.510	0.126
9	What to do if a stranger approaches you	0.765	0.424	-1.628	0.073	-1.791	0.101
10	Taking care of your teeth	0.765	0.424	-1.626	0.072	-1.832	0.097
11	What to do if someone chokes on food	0.762	0.426	-1.604	0.071	-1.699	0.096
12	Safety at home, school, or play	0.828	0.378	-2.157	0.073	-2.299	0.095
13	Stress	0.641	0.480	-0.789	0.067	-0.858	0.091
14	How to handle conflict	0.779	0.415	-1.733	0.072	-1.827	0.099
15	Where to go for help with a health problem	0.826	0.380	-2.137	0.073	-2.217	0.092
16	The problems of being underweight	0.554	0.497	-0.268	0.066	-0.291	0.084
17	Suicide	0.681	0.466	-1.044	0.069	-1.055	0.100

The estimates for σ_ξ^2 of the 17 items in the MRM-CS are listed in the left panel of Table 8. It appeared that LPD was significant for most items, suggesting that the acquiring of general knowledge in daily life was related to the sampled communities. The two items with the largest LPD were "AIDS" ($\sigma_\xi^2 = 0.743$) and "What to do if a stranger approaches you" ($\sigma_\xi^2 = 0.594$). It seems that these two topics were stressed, to different extents, in different regions. The two items with the smallest LPD were "The problems of being overweight" ($\sigma_\xi^2 = 0.171$) and "The problems of being underweight" ($\sigma_\xi^2 = 0.108$), suggesting that all sampled communities had similar degrees of highlighting the importance of weight control. To illustrate the presence of LPD, the MRM-CS was fit to a replicated dataset in which students were randomly grouped into schools. The estimates of σ_ξ^2 for the random dataset (listed in the right panel of Table 8) ranged from 0.031 to 0.086, which were much smaller than those for the real dataset. Thus, the residual dependence among students in real schools was not due to random errors.

The σ_ε^2 and σ_u^2 estimates under the MRM were 0.413 and 2.198, respectively, whereas the estimates under the MRM-CS were 0.452 and 2.431, respectively, suggesting that the MRM-CS yielded a wider distribution of the latent trait. On the other hand, a narrower range of the item parameters was found under the MRM, suggesting that ignoring LPD resulted in a shrunken scale. The ICC under the MRM and the MRM-CS was 0.158 and 0.157, respectively. Consistent with the findings in the simulation studies, the ICC under the MRM was slightly larger than the ICC under the MRM-CS because moderate influences of LPD were found. In addition, these two models yielded a similar measurement precision: The test reliability of the survey was .780 in the MRM and .781 in the MRM-CS.

Table 8. LPD estimates for real schools and randomly grouped schools under the MRM-CS in the general knowledge scale.

		Real schools		Random schools	
No.	Item	Estimate	SE	Estimate	SE
1	The foods you should and shouldn't eat	0.248	0.074	0.053	0.023
2	The importance of exercise	0.277	0.095	0.052	0.024
3	Smoking	0.273	0.089	0.084	0.039
4	The problems of being overweight	0.171	0.045	0.031	0.011
5	Drinking	0.370	0.116	0.086	0.042
6	Drug abuse	0.334	0.124	0.072	0.038
7	Pregnancy	0.553	0.116	0.059	0.025
8	AIDS	0.743	0.161	0.080	0.036
9	What to do if a stranger approaches you	0.594	0.110	0.057	0.023
10	Taking care of your teeth	0.558	0.100	0.043	0.017
11	What to do if someone chokes on food	0.390	0.080	0.045	0.017
12	Safety at home, school, or play	0.287	0.073	0.079	0.030
13	Stress	0.301	0.061	0.037	0.013
14	How to handle conflict	0.357	0.085	0.051	0.019
15	Where to go for help with a health problem	0.232	0.064	0.050	0.022
16	The problems of being underweight	0.108	0.037	0.032	0.013
17	Suicide	0.458	0.090	0.051	0.019

Two pairs of schools with mean scores of 11.2 and 12.27 were selected to illustrate the performance of HCA. Figures 6 and 7 show the dendrograms of the hierarchical clustering of person residuals for the two school pairs under the two models. As shown in Figure 6a, when the residuals under the MRM were analyzed, the two explored clusters approximated the two real schools, implying that the residuals of adolescents from the same school were

homogeneous. Conversely, when the residuals under the MRM-CS were analyzed, as shown in Figure 6b, the correspondence between the explored clusters and the real schools was not clear. As shown in Figure 7a, the classification was not as distinct as the pattern in Figure 6a. Comparing Figures 7a and b, one finds that the two-class solutions under the two models were very different. The Euclidean distance for the good-classification pair (Figure 6) was 8.20, but was 5.66 for the poor-classification pair (Figure 7), which supports the inference that higher heterogeneous person residuals between clusters (i.e., a larger Euclidean distance) can improve the performance of HCA in detecting LPD. In sum, the findings of HCA indicated that the magnitudes of LPD were substantial and should not be ignored.

Furthermore, both the MRM and MRM-CS were fit as a survey of general knowledge to examine the influence of LPD. Noticeable contextual effects were found in some items with moderate magnitude. The consequence of ignoring LPD by fitting the MRM led to a shrunken scale, a higher ICC, and a lower test reliability.

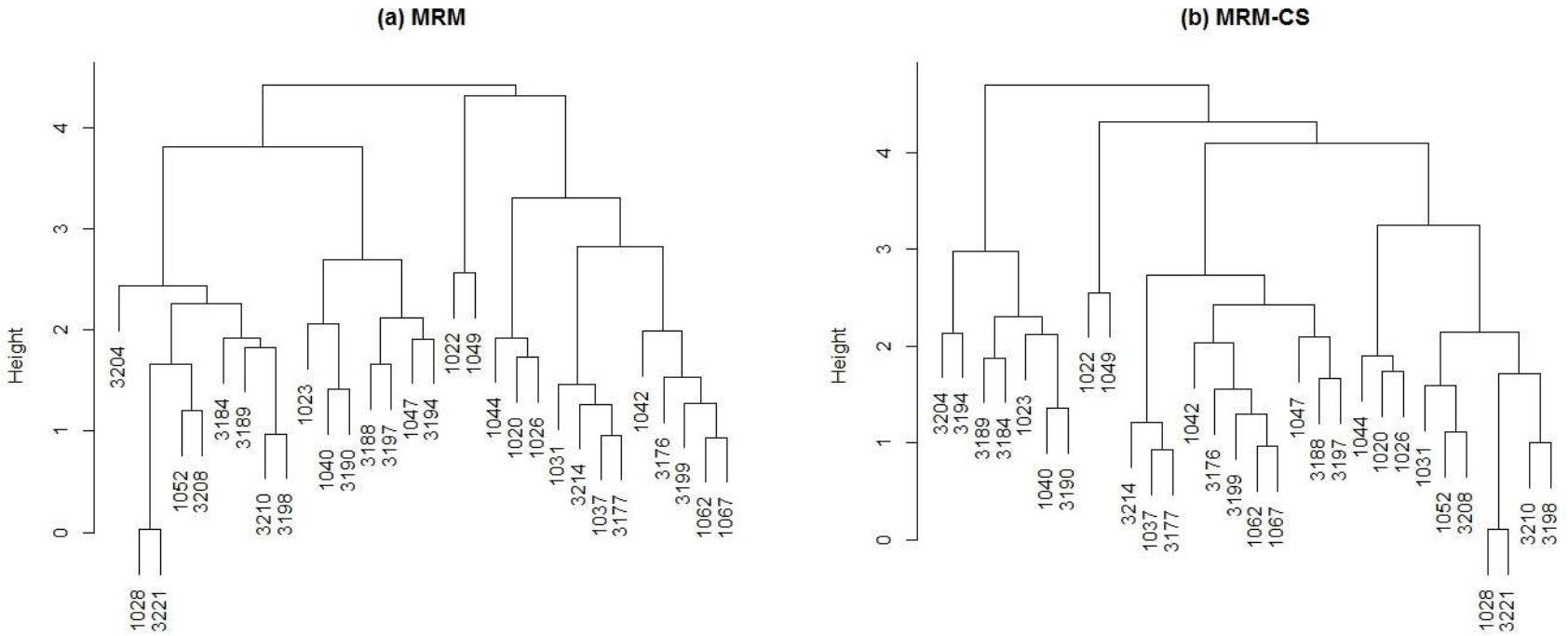

Figure 6. Dendrograms for subjects nested within two schools (M = 12.5) in the general knowledge scale.

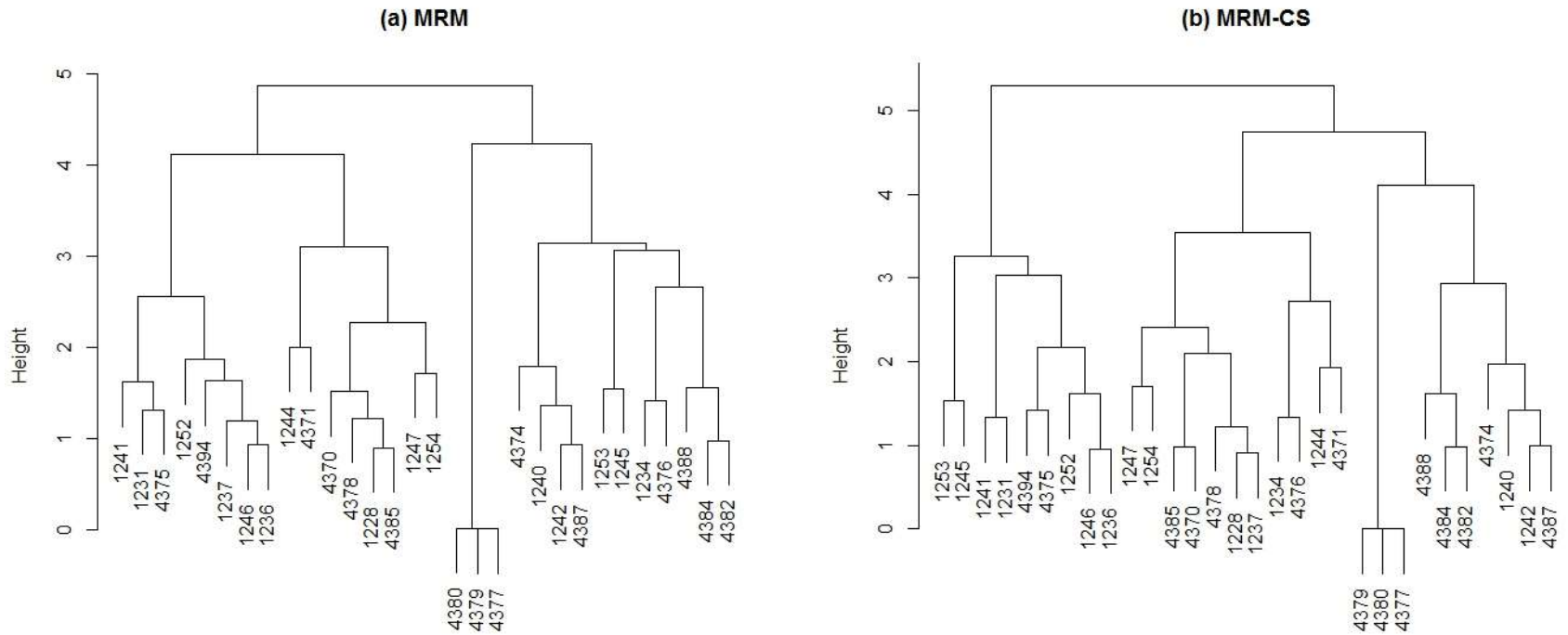

Figure 7. Dendrograms for subjects nested within two schools (M = 13.0) in the general knowledge scale.

5.2 Example 2: Problem-solving capability of police officers

The data used for the second example are from the Impact of Community Policing Training and Program Implementation on Police Personnel in Arizona project (Haarr, 2003), a longitudinal study examining the impact of the Phoenix Regional Training Academy's curriculum on police trainees. This instrument, which was designed to measure police officers' attitudes and beliefs in several aspects of the job, was administered on four occasions. The study used a dataset of 444 police officers responding to five four-point Likert-scale items (see Table 9) regarding problem-solving capabilities. Both the MPCM and MPCM-CS were fit to the data. Because of the repeated measures design and the usage of common items across time, substantial GPD and LPD were expected. Thus, fitting the MPCM-CS is effective for quantifying both kinds of dependences. Furthermore, HCA was conducted to investigate the homogeneity of cross-time person residuals from the MPCM- and MPCM-CS. Police officers were grouped based on the mean scores across four administrations.

Table 9. LPD estimates for real and randomized repeated measures under the MPCM-CS in the problem-solving scale.

Item	Real Repeated Measures		Randomized Repeated Measures	
	$\hat{\sigma}^2_\xi$	SE	$\hat{\sigma}^2_\xi$	SE
1. Identify community problems	1.411	0.388	0.264	0.163
2. Use problem-solving techniques to analyze problems	2.126	0.461	0.496	0.277
3. Develop solutions to community problems	0.259	0.157	0.069	0.036
4. Evaluate solutions to see how well they work	0.872	0.300	0.110	0.062
5. Work with beat residents to solve problems in the neighborhood	1.610	0.416	0.122	0.084

Figure 8 displays the threshold estimates under the MPCM and the MPCM-CS, showing noticeable discrepancies between two sets of estimates. The difference between two corresponding parameter estimates (in absolute value) was between 0.270 and 1.540, suggesting that police officers exhibited substantial LPD across four administrations on these five items.

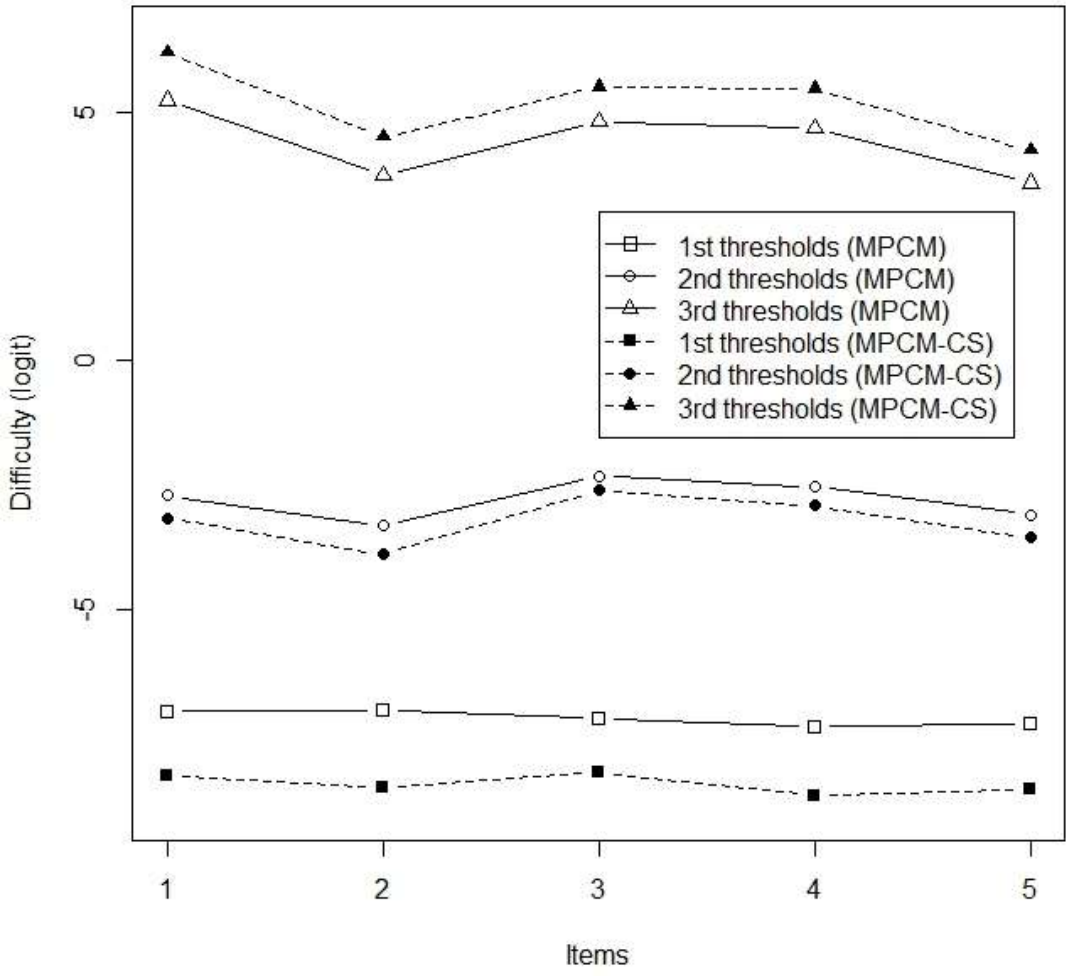

Figure 8. Threshold estimates under the MPCM and the MPCM-CS in the problem-solving scale.

The AIC, BIC, and DIC for the MPCM were 6.684, 6,774, and 7,720, respectively, and

for the MPCM-CS, they were 5,724, 5,840, and 7,225, respectively, suggesting that the MPCM-CS had a better fit. The posterior predictive *p*-value for the MPCM and the MPCM-CS were 0.888 and 0.800, respectively. The expected mean and standard deviation of Q_3 would be -0.0007 and 0.577, respectively. The empirical mean Q_3 was 0.174 for the MPCM and 0.094 for the MPCM-CS, and the empirical standard deviation was 0.619 for the MPCM and 0.542 for the MPCM-CS, which suggests that the MPCM-CS had a slightly better fit than did the MPCM.

The estimates for σ_ξ^2 of the five items in the MPCM-CS are listed in the left panel of Table 9. The item "Use problem-solving techniques to analyze problems" had the largest variance ($\sigma_\xi^2 = 2.126$). In addition, the MPCM-CS was fit to a replicated dataset in which the repeated measures were randomized to illustrate the presence of LPD. The estimates of σ_ξ^2 for the randomized repeated measures (listed in the right panel of Table 9) ranged from 0.069 to 0.496, which were much smaller than those for the real dataset. The residual dependence among repeated measures in the real dataset was confirmed accordingly.

The inter- and intra-person variances under the MPCM were 3.527 and 3.987, respectively, and 4.735 and 5.629, respectively, under the MPCM-CS. The ICC was .469 in the MPCM and 0.457 in the MPCM-CS, suggesting the consistency of proficiency across time. Comparing the inter- and intra-person variance estimates with the LPD estimates, trivial influences of LPD on the test reliabilities were expected. The test reliability was .793 in the MPCM and .788 in the MPCM-CS.

HCA was applied to analyze the problem-solving scale. Unfortunately, the produced dendrograms based on the residuals under the MPCM and MPCM-CS were almost identical because of the short test length. The results implied that HCA may be unfeasible for detecting LPD in short tests.

5.3 Example 3: Student surveys in the ICCS 2009

The International Civic and Citizenship Study (ICCS) in 2009 (Schultz, Ainley, & Frailon, 2011), which surveyed 14-year-old students in 28 countries, was used in the third example. Specifically, a set of subscales that included 17 four-point Likert scale items (i.e., 0 = *not at all*, 1 = *to a small extent*, 2 = *to a moderate extent*, and 3 = *to a large extent*) was designed to assess students' perceptions of school contexts, including open classrooms (six items), student influence (six items), and student-teacher relations (five items). Appendix E lists the items on these three subscales. A Taiwan sample of 5,006 students from 150 schools (150 classes, to be precise) was selected. Each school included 20 to 51 students. Both the MDMPCM and MDMPCM-CS were fit to the data, and the genders of students were dummy-coded (boys = 1 and girls = -1) to predict the three latent traits. It was noticed that some items were about the learning environments (e.g., schools, classes, and teachers) and thus might be distinctive across classes. Therefore, we examined the hypothesis that students sampled from the same class would be more homogeneous than others from different classes in their perceptions. Because a teacher's leadership might influence the general mood of a class and the student-teacher relationship (Brophy, 2006), a teacher's influence would be revealed from students' homogeneity in their overall perceptions (i.e., GPD) measured by the three subscales and some specific item descriptions (i.e., LPD). It was anticipated that the relationships among the three perceptions could be captured by means of multi-dimensional modeling. Fitting the MDMPCM-CS would be efficient for uncovering the similarities among the three perceptions and the extent of the homogeneity

on person residuals by items simultaneously. Ignoring LPD by fitting the simpler MDMPCM, according to the findings from the simulation study, would yield shrunken scales and attenuated factor loadings when the magnitudes of LPD were substantial. Likewise, person residuals under the two models were used to conduct HCA. Schools were paired according to the mean scores by each scale.

Figure 9 displays the threshold estimates under the MDMPCM and the MDMPCM-CS, indicating that these estimates roughly overlapped, with minor discrepancies. Among the three subscales, the difference between two corresponding parameter estimates (in absolute value) was between 0.001 and 0.136, implying that students exhibited LPD on only some items and that the influence was trivial.

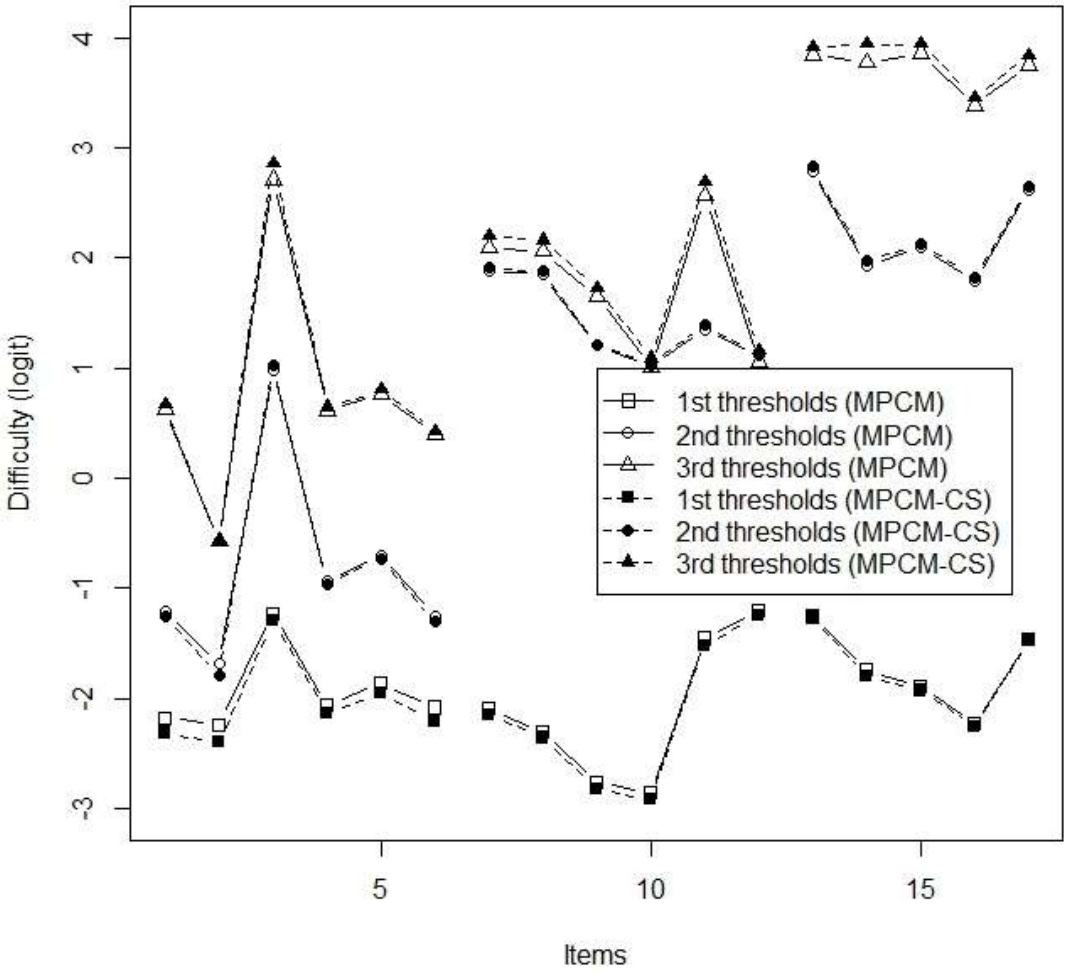

Figure 9. Threshold estimates under the MDMPCM and the MDMPCM-CS in the ICCS survey.

The AIC, BIC, and DIC for the MDMPCM were 146,512, 146,942, and 158,219, respectively, and for the MDMPCM-CS, they were 144,130, 144,671, and 156,800, respectively, suggesting that the MDMPCM-CS had a better fit. The posterior predictive *p*-values for the MDMPCM and the MDMPCM-CS were 0.005 and 0.990, respectively, indicating that the MDMPCM yielded an underfit, whereas the MDMPCM-CS yielded an overfit. Table 10 lists the empirical Q_3 values for the individual subscales and the whole

scale. Overall, the empirical mean Q_3 was 0.001 for the MDMPCM and 0.001 for the MDMPCM-CS, and the empirical standard deviation was 0.285 for the MDMPCM and 0.283 for the MDMPCM-CS.

Table 10. Expected and empirical Q_3 under the MDMPCM and MDMPCM-CS in the ICCS survey.

	No. of	No. of	Expected Q_3		MDMPCM		MDMPCM-CS	
Subscale	Items	Students	Mean	SD	Mean	SD	Mean	SD
OC	6	5,006	-0.0002	0.500	0.002	0.468	0.002	0.466
SI	6	5,006	-0.0002	0.500	0.042	0.524	0.032	0.510
STR	5	5,006	-0.0002	0.577	0.052	0.546	0.044	0.533
Total	17	5,006	-0.0002	0.258	0.001	0.285	0.001	0.283

Note. OC = open classrooms, SI = student influence, and STR = student-teacher relations.

The estimates for σ^2_ξ for the 17 items in the MDMPCM-CS are listed in the left panel of Table 11. It appears that LPD was significant for a few items. In the open-classrooms scale, the item "Students bring up current political events for discussion in class" had the largest variance ($\sigma^2_\xi = 0.290$). On the student-influence subscale, the item "Classroom rules" had the largest variance ($\sigma^2_\xi = 0.252$). On the student-teacher relation scale, the item "Students get along well with most teachers" had the largest variance ($\sigma^2_\xi = 0.150$). Noticeably, the large dependence in these items was reasonable because they related to teachers' class management. To illustrate the presence of LPD, the MDMPCM-CS was fit to a replicated dataset in which students were randomly grouped into schools. The estimates of σ^2_ξ for the random dataset (listed in the right panel of Table 11) ranged from 0.018 to 0.095, a smaller range than those for the real dataset. Thus, the residual dependence among students in the real dataset was not due to random errors.

Table 11. LPD estimates for real schools and randomly grouped schools under the MDMPCM-CS in the ICCS survey.

Subscale	Item	Real Schools		Random Schools	
		$\hat{\sigma}^2_\xi$	SE	$\hat{\sigma}^2_\xi$	SE
Open Classrooms	IS2G16B	0.042	0.012	0.023	0.007
	IS2G16C	0.047	0.015	0.021	0.006
	IS2G16D	0.290	0.047	0.053	0.014
	IS2G16E	0.112	0.021	0.030	0.010
	IS2G16F	0.022	0.007	0.018	0.005
	IS2G16G	0.036	0.011	0.020	0.006
Student Influence	IS2G17A	0.050	0.018	0.024	0.007
	IS2G17B	0.023	0.008	0.021	0.007
	IS2G17C	0.025	0.008	0.018	0.006
	IS2G17D	0.065	0.017	0.020	0.007
	IS2G17E	0.252	0.042	0.095	0.022
	IS2G17F	0.091	0.020	0.025	0.007
Student -Teacher Relations	IS2G18A	0.067	0.023	0.045	0.016
	IS2G18B	0.150	0.033	0.035	0.013
	IS2G18C	0.042	0.015	0.029	0.010
	IS2G18E	0.043	0.015	0.029	0.010
	IS2G18F	0.065	0.022	0.044	0.016

Note. Full item contents can be found in Appendix E.

The parameter estimates under the MDMPCM and MDMPCM-CS are shown in Table 12. The variances for these three latent traits were 1.357, 2.553, and 4.094, respectively, under the MDMPCM, and 1.522, 2.746, and 4.240, respectively, under the MDMPCM-CS. Comparing the two sets of estimates, the scale shrinkage in the MDMPCM was not very serious because the magnitudes of LPD were minor compared with the variances of the latent traits. In the MDMPCM-CS, the correlation matrix for the individual-level effects was $\begin{bmatrix} 1.000 & -0.239 & -0.398 \\ -0.239 & 1.000 & 0.327 \\ -0.398 & 0.327 & 1.000 \end{bmatrix}$, and $\begin{bmatrix} 1.000 & -0.336 & -0.667 \\ -0.336 & 1.000 & 0.536 \\ -0.667 & 0.536 & 1.000 \end{bmatrix}$ for the school-level effects, suggesting that, at both the student and school levels, the influence score and

student-teacher relations score were positively correlated, but the open-classroom score was negatively correlated with the influence score and the student-teacher relation score. The results also indicate that boys had a higher mean score in evaluating the two subscales of open schools and student influence, whereas girls had a higher mean score in evaluating student-teacher relationships. The influences of LPD on test reliabilities were trivial. The test reliabilities of the three subscales in the MDMPCM were .798, .847, and .845, respectively, and .803, .848, and .844, respectively, for the MDMPCM-CS.

Table 12. Multilevel modeling results under the MDMPCM and MDMPCM-CS.

	MDMPCM		MDMPCM-CS	
Parameters	Estimate	SE	Estimate	SE
Gender (OC)	0.195	0.018	0.186	0.018
Gender (SI)	0.101	0.024	0.100	0.024
Gender (STI)	-0.184	0.029	-0.179	0.029
$\sigma^2_{\varepsilon 11}$	1.305	0.045	1.183	0.040
$\sigma^2_{\varepsilon 12}$	-0.441	0.034	-0.410	0.031
$\sigma^2_{\varepsilon 13}$	-0.900	0.044	-0.845	0.042
$\sigma^2_{\varepsilon 22}$	2.614	0.078	2.428	0.078
$\sigma^2_{\varepsilon 23}$	1.048	0.059	0.997	0.057
$\sigma^2_{\varepsilon 33}$	3.924	0.125	3.763	0.136
$\sigma^2_{u 11}$	0.191	0.030	0.149	0.023
$\sigma^2_{u 12}$	-0.055	0.020	-0.045	0.018
$\sigma^2_{u 13}$	-0.160	0.033	-0.129	0.027
$\sigma^2_{u 22}$	0.142	0.026	0.131	0.025
$\sigma^2_{u 23}$	0.111	0.029	0.109	0.030
$\sigma^2_{u 33}$	0.302	0.052	0.309	0.053

Note. OC = open classrooms, SI = student influence, and STR = student-teacher relations.

The raw scores of the open class scale were used herein to conduct HCA because large LPD was observed on the open classroom scale. For simplicity, the results for two pairs of schools with mean scores of 11.2 and 12.27 are presented. Figures 10 and 11 show the dendrograms of the hierarchical clustering of person residuals for the two school pairs under the MDMPCM and MDMPCM-CS, respectively. As shown in Figure 10a, when the residuals under the MDMPCM were analyzed, the two explored clusters approximated the two real schools, implying that the residuals of adolescents from the same school were homogeneous. Conversely, when the residuals under the MDMPCM-CS were analyzed, as shown in Figure 10b, the correspondence between the explored clusters and the real schools was not clear. In another school pair (see Figure 11), the correspondence between the explored clusters and the real schools was almost invisible. The results of HCA generally indicated minor LPD in the ICCS dataset.

Overall, the LPD in the ICCS dataset can be identified by the MDMPCM-CS. Students tended to respond homogeneously on a few items about teachers' class management. Nevertheless, the LPD in this example was not very serious because the heterogeneity of the latent traits across students was much larger than that in the students' residuals.

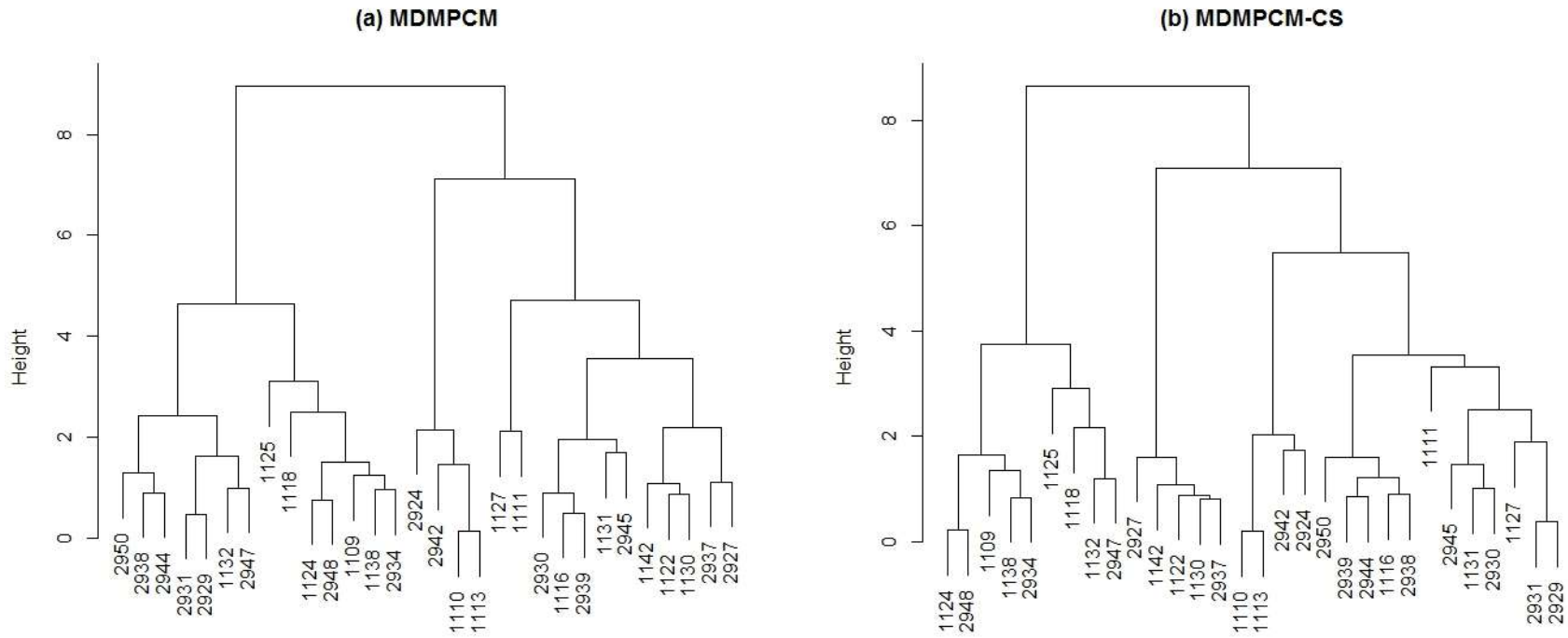

Figure 10. Dendrograms for subjects nested within two schools (M = 11.2) on the open classroom scale.

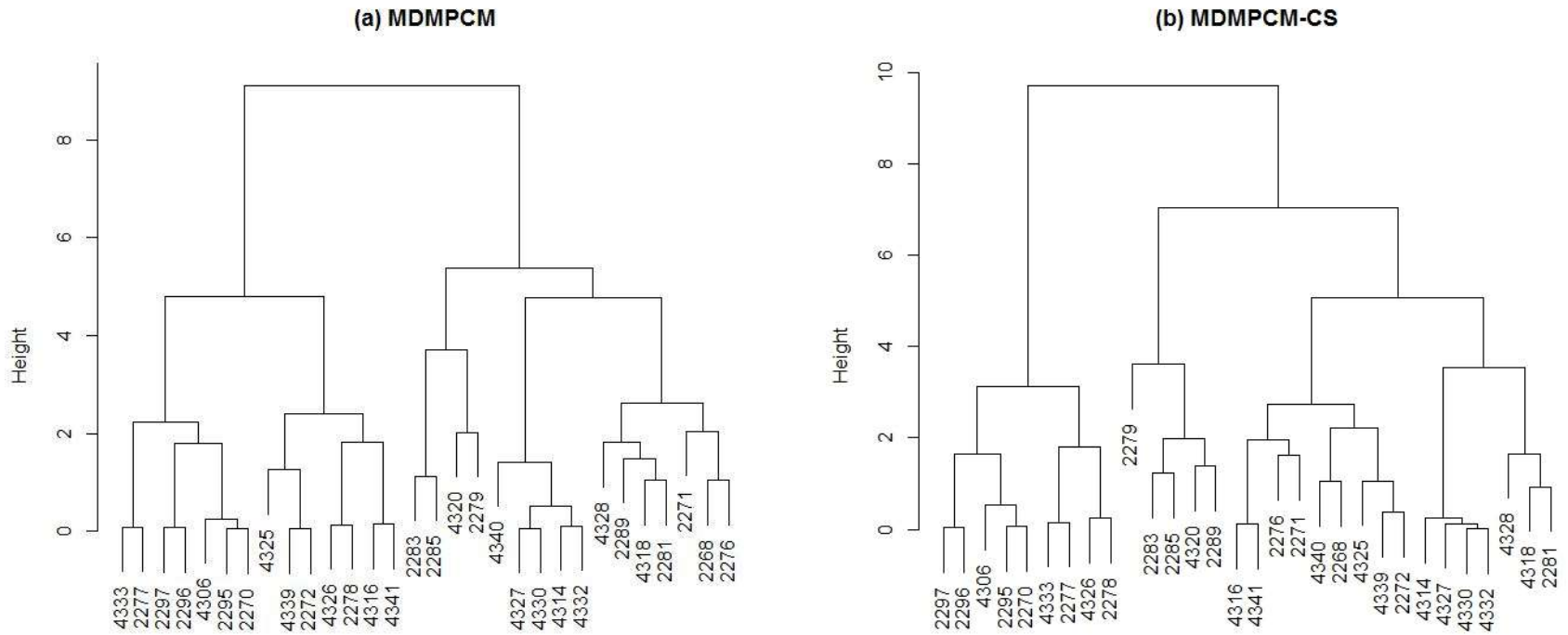

Figure 11. Dendrograms for subjects nested within two schools (M = 12.27) on the open classroom scale.

5.4 Example 4: Love Relationship Scale

The aforementioned paired-sample design is a special application of cluster sampling, so a survey with paired samples was selected as the fourth example. The love relationship scale (C. F. Wang, 2000), which was developed based on the triangular theory of love (Sternberg, 1986, 1987), consists of three subscales (eight items in each) measuring passion, intimacy, and commitment. A common six-point rating scale was designed for these three subscales: 0 = *very inconsistent*, 1 = *quite inconsistent*, 2 = *inconsistent*, 3 = *consistent*, 4 = *quite consistent*, and 5 = *very consistent*. Therefore, a respondent who had high scores on these three subscales exhibited higher passion, intimacy, and commitment in his (or her) relationship. The reader can refer to Appendix F for the item descriptions of all items (in Chinese). The survey included 202 couples in Taiwan. According to Sternberg (1986), the relationship between the general love perception and the three specific love components follows a higher-order structure so that the HOPCM and HOPCM-CS are applicable. In particular, because few respondents chose the extreme-response categories on some items, a set of common threshold parameters was used for items within the same subscale, as constraints in the RSM, to ease the burden of estimation for the threshold parameters. HCA also was applied to examine whether the original pairing structure could be recovered by means of the information of person residuals from the HOPCM and HOPCM-CS.

Table 13 summarizes the mean and standard deviation of the raw scores and the mean thresholds under the HOPCM and HOPCM-CS for each item. According to the distribution of the raw scores, respondents' responses were generally positive, suggesting that they were satisfied with their relationships. The respondents showed higher agreement on the passion and intimacy subscales than on the commandment scale. In terms of the mean threshold estimates, it was evident that the HOPCM yielded a narrower range on each scale than did the HOPCM-CS, indicating substantial LPD.

The AIC, BIC, and DIC for the HOPCM were 19,769, 19,945, and 20,618, respectively, and for the HOPCM-CS, they were 18,172, 18,444, and 19,970, respectively. These indices indicated that the HOPCM-CS had a better fit than did the HOPCM. The posterior predictive *p*-values for the HOPCM and the HOPCM-CS were 0.080 and 0.378, respectively. Table 14 lists the empirical Q_3 values for the individual subscales and the whole scale. Overall, the empirical mean Q_3 was -0.002 for the HOPCM and -0.002 for the HOPCM-CS, and the empirical standard deviation was 0.231 for the HOPCM and 0.228 for the HOPCM-CS.

Table 13. Raw scores and mean threshold estimates under the HOPCM and HOPCM-CS on the Love Relationship Scale.

		Raw score		HOPCM		HOPCM-CS	
Subscale	Item	Mean	SD	Estimate	SE	Estimate	SE
Passion	2	3.938	0.886	-2.507	0.149	-2.981	0.189
	4	3.940	0.795	-2.517	0.157	-2.974	0.188
	5	3.774	0.907	-2.108	0.148	-2.538	0.194
	9	4.196	0.834	-3.189	0.159	-3.850	0.224
	11	3.532	0.878	-1.555	0.147	-1.843	0.165
	19	4.161	0.782	-3.093	0.159	-3.649	0.199
	23	3.819	0.911	-2.207	0.147	-2.610	0.185
	24	3.489	1.042	-1.458	0.142	-1.726	0.164
Intimacy	1	4.122	0.740	-2.240	0.125	-2.872	0.162
	3	4.035	0.929	-2.058	0.124	-2.689	0.167
	6	3.814	0.835	-1.624	0.113	-2.100	0.150
	8	3.715	0.847	-1.449	0.110	-1.879	0.143
	15	3.676	0.851	-1.381	0.113	-1.795	0.139
	16	4.490	0.884	-3.172	0.141	-4.690	0.258
	18	3.579	1.200	-1.222	0.106	-1.699	0.158
	22	3.923	1.041	-1.826	0.121	-2.472	0.168
Commitment	7	3.822	1.022	-1.875	0.121	-2.170	0.142
	10	3.195	1.224	-0.907	0.104	-1.038	0.133
	12	3.970	1.047	-2.146	0.122	-2.485	0.151
	13	3.162	1.325	-0.860	0.110	-1.007	0.127
	14	3.935	1.094	-2.082	0.120	-2.438	0.159
	17	3.365	1.300	-1.153	0.113	-1.337	0.134
	20	2.915	1.488	-0.520	0.103	-0.620	0.133
	21	3.239	1.311	-0.969	0.108	-1.128	0.130

Note. Raw scores of items 10, 12, 13, 14, 17, 20, and 21 have been reversely coded.

Table 14. Expected and empirical Q_3 under the HOPCM and HOPCM-CS on the Love Relationship Scale.

Subscale	No. of Items	No. of Students	Expected Q_3		HOPCM		HOPCM-CS	
			Mean	SD	Mean	SD	Mean	SD
Passion	8	404	-0.002	0.408	0.003	0.402	0.004	0.397
Intimacy	8	404	-0.002	0.408	0.000	0.396	0.002	0.391
Commitment	8	404	-0.002	0.408	0.005	0.407	0.006	0.398
Total	24	404	-0.002	0.213	-0.002	0.231	-0.001	0.228

The estimates for σ_ξ^2 of the 24 items in the HOPCM-CS are listed in the left panel of Table 15. It appears that LPD was substantial for some items. For example, on the passion scale, Item 9, "I hope he or she would think I'm attractive" has the largest variance (σ_ξ^2 = 1.941). On the intimacy subscale, Item 16, "We have intimate act of kissing" has the largest variance (σ_ξ^2 = 4.259), and on the commitment scale, Item 14, "Sometimes I feel like I am not sincere with him (or her) (negatively worded)" has the largest variance (σ_ξ^2 = 0.617). To illustrate the presence of LPD, the HOPCM-CS was fit to a replicated dataset in which couples were randomly paired. The estimates of σ_ξ^2 for the artificial dataset (listed in the right panel of Table 15) were generally smaller than those of the real dataset, but some estimates were not close to zero. The large dependence in these items might be because these descriptions are relative to the general thoughts or behaviors in the physical-attraction stage of mate selection. In sum, the residual dependence between the real couples was not due to random errors.

Table 15. LPD estimates for real couples and randomly grouped couples under the HOPCM-CS on the Love Relationship Scale.

Subscale	Item	Real couples		Random couples	
		$\hat{\sigma}^2_\xi$	SE	$\hat{\sigma}^2_\xi$	SE
Passion	2	0.503	0.217	0.289	0.178
	4	0.083	0.052	0.086	0.049
	5	1.112	0.323	0.591	0.241
	9	1.941	0.460	1.919	0.458
	11	0.231	0.128	0.231	0.141
	19	0.108	0.071	0.110	0.077
	23	0.142	0.100	0.149	0.086
	24	0.145	0.095	0.128	0.067
Intimacy	1	0.087	0.048	0.083	0.048
	3	0.294	0.162	0.209	0.127
	6	0.114	0.082	0.125	0.075
	8	0.073	0.042	0.062	0.035
	15	0.084	0.054	0.064	0.034
	16	4.529	0.968	2.643	0.604
	18	1.198	0.272	1.302	0.317
	22	0.852	0.248	0.635	0.228
Commitment	7	0.082	0.046	0.065	0.032
	10	0.086	0.046	0.083	0.045
	12	0.252	0.139	0.138	0.089
	13	0.272	0.133	0.114	0.064
	14	0.617	0.203	0.187	0.108
	17	0.143	0.086	0.107	0.062
	20	0.589	0.181	0.244	0.106
	21	0.259	0.126	0.163	0.088

The parameter estimates in the hierarchical structure are listed in Table 16. The standardized estimates for the three loadings under the HOPCM were .842, .892, and .833, respectively, whereas the estimates under the HOPCM-CS were .864, .877, and .835, respectively. In treating the HOPCM-CS as the gold standard, it was suggested that ignoring

the substantial LPD would decrease the relationships between the general love perception and three specific love components. Because the LPD was substantial on most items, other statistics -- such as the test reliabilities, gender difference, and clustering effect in the general love perception -- varied between the two models. The test reliabilities for the three first-order and one second-order latent traits in the HOPCM were .892, .869, .917, and .847, respectively, and .900, .873, .917, and .848, respectively, in the HOPCM-CS. The gender difference in the general love perception was 0.109 in the HOPCM and 0.133 in the HOPCM-CS, indicating that females exhibited a higher involvement in love than did their companions. The conditional ICC on the second-order latent trait was .487 in the HOPCM and 0.459 in the HOPCM-CS, indicating that the couples exhibited a high degree of general love perception, and LPD would result in inflation when fitting the standard model.

Table 16. Parameter estimates under the HOPCM and HOPCM-CS on the Love Relationship Scale.

	HOPCM		HOPCM-CS	
Parameters	Estimate	SE	Estimate	SE
λ_1 (Passion)	1.000*	–	1.000*	–
λ_2 (Intimacy)	0.790	0.067	0.784	0.063
λ_3 (Commitment)	0.974	0.073	0.914	0.082
ν_1 (Passion)	0.712	0.119	0.912	0.172
ν_2 (Intimacy)	0.273	0.072	0.486	0.107
ν_3 (Commitment)	0.721	0.112	0.956	0.158
Gender	-0.109	0.054	-0.133	0.067
σ^2_ε	0.894	0.159	1.431	0.252
σ^2_u	0.849	0.181	1.211	0.267

Note. * = constrained for identification.

HCA was applied to analyze the Love Relationship Scale. Because the most salient LPD was observed on items 16 and 18 on the intimacy scale, the couples were stratified according to the mean score on the intimacy scale for the subsequent HCA. Two strata with

mean scores of 29.5 and 31.5 are presented as examples, and the dendrograms are shown in Figures 12 and 13. When the residuals under the HOPCM were analyzed, as shown in Figures 12a and 13a, males and females were grouped in a disorderly fashion, and the original pairing structure was nearly unidentified. Thus, it was not surprising that the original pairing structure also was unobserved when the residuals under the HOPCM-CS were analyzed, as shown in Figures 12b and 13b. The results pointed out an evident limitation of HCA for paired samples.

On the Love Relationship Scale, a clear conclusion can be stated that the paired males and females yielded not only similar attitudes on their relationships, but also common understandings toward some interactive behaviors, and the HOPCM-CS successfully detached such dependence from the paired couples. Because the influences of common views toward some aspects were not identified in the standard HOPCM, the similarities in attitudes toward relationships would be overestimated.

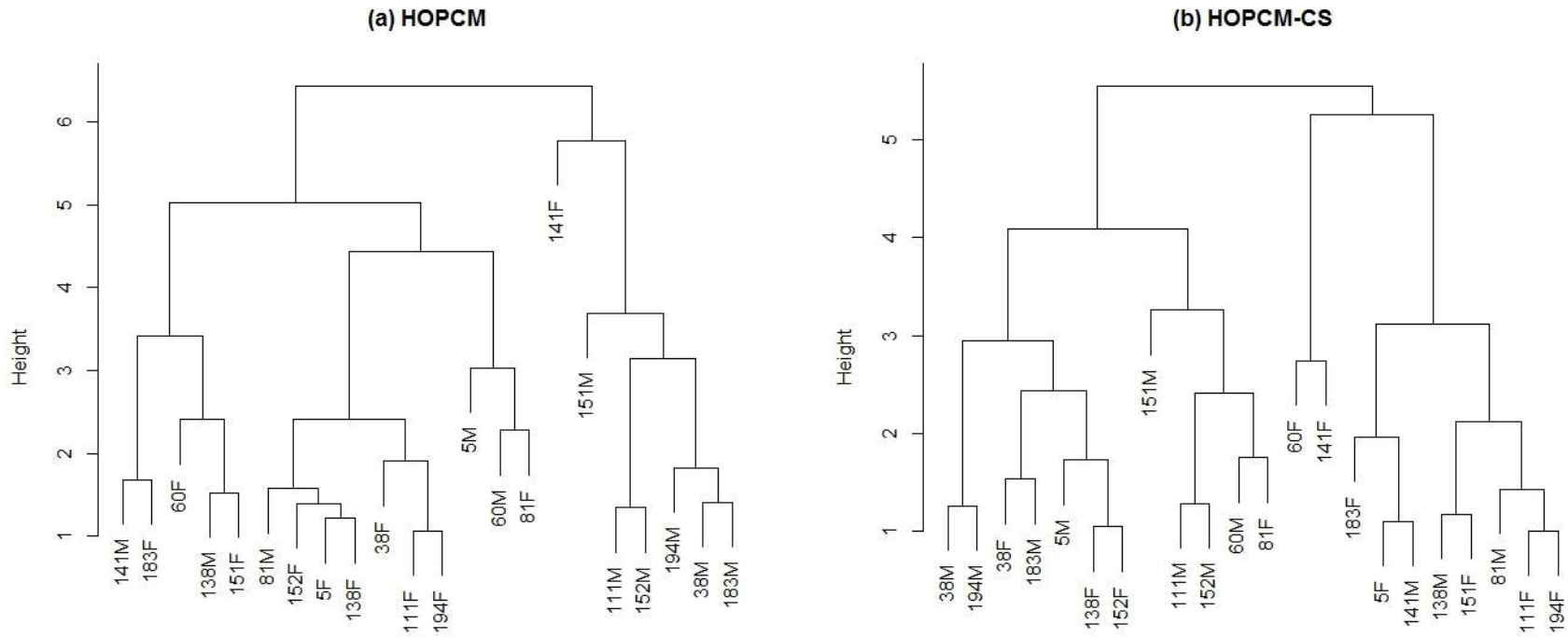

Figure 12. Dendrograms for couples that had a mean score of 29.5 on the intimacy scale.

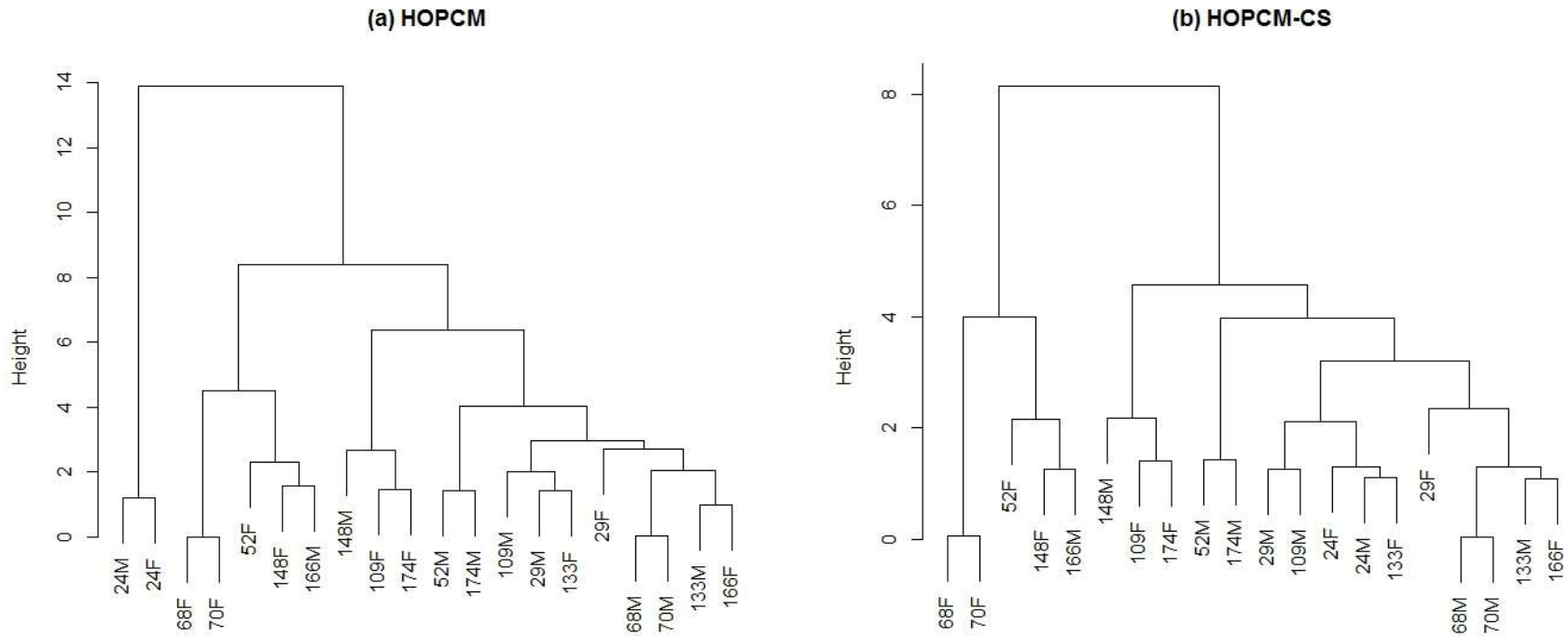

Figure 13. Dendrograms for couples that had a mean score of 31.5 on the intimacy scale.

Chapter 6: Discussion and Conclusions

6.1 Summary

When two- or multiple-stage sampling is conducted, persons nested within a cluster may perform more similarly than those from different clusters (i.e., GPD), and there may be local dependence among persons in the same cluster after standard multilevel IRT models are fit (i.e., LPD). The traditional approach of multilevel modeling in IRT is aimed at quantifying the magnitude of GPD, but it ignores the possibility of LPD. In addition, although LID and LPD are two major violations of the local independence assumption, the investigation of LPD is scarce in the IRT literature.

To answer the first research question, relevant literature in relation to multilevel modeling and violation of local independence are reviewed. The MRM-CS and MPCM-CS—in which multilevel modeling is created on the intended-to-be-measured latent trait to account for GPD, and a set of random variables is implemented to quantify LPD in different items—are proposed for dichotomous and polytomous responses. To answer the second research question, two multidimensional generalizations of the MDMPCM-CS and HOPCM-CS are then proposed for tests consisting of multiple subtests. Compared with the MDMPCM-CS, the HOPCM-CS is especially suitable for the hierarchy structures of latent traits measured by different subtests. In addition, the MFRM-CS is presented to explain GPD and LPD in multifaceted data. To avoid highly dimensional integration in parameter estimation for these new models, MCMC methods are adopted, and the freeware program WinBUGS is used for parameter estimation.

To evaluate the applicability of the new models and answer the third research question, several simulation studies were conducted to evaluate the parameter recovery of the new models and the consequences of failing to consider LPD. The WinBUGS codes in the simulations are attached in Appendices A to D. Simulation study 1 examined the parameter

recovery of the MPCM-CS. It was found that, when item responses were generated from the MPCM-CS, the parameters could be recovered fairly well, and fitting the standard MPCM resulted in biased parameter estimates, shrunken scales, inflated GPD, and deflated test reliabilities. In contrast, when item responses were generated from the MPCM, fitting the unnecessarily complicated MPCM-CS yielded results very similar to those obtained by fitting the true model, and estimates of the LPD parameters were approximately zero. Simulation study 2 investigated the influence of different combinations of clusters and intra-cluster sample sizes, given a fixed total sample size. It concluded that the larger the number of clusters, the more accurate the parameter estimation would be. Simulation studies 3 and 4 examined the parameter recovery of the MDMPCM-CS and the HOPCM-CS. The results were similar to the findings in simulation study 1. Simulation study 5 examined the MFRM-CS and confirmed its feasibility for multifaceted data.

To answer the fourth research question, a conventional approach for HCA was adopted in this study to evaluate possible LPD. The concept is intuitive: When a standard model is fit to data with LPD, the resulting person residuals scatter in a multidimensional space, so HCA can help recover the cluster membership with person residuals. Simulations were conducted to evaluate the performance of HCA in detecting LPD, but the findings indicated its limited value compared with direct IRT modeling to LPD.

To answer the last research question, four empirical examples – one subscale in the National Longitudinal Study of Adolescent Health (Add Health) project measuring the general knowledge in daily life, one subscale in the longitudinal Impact of Community Policing Training and Program Implementation on Police Personnel in Arizona study, a subset of a student questionnaire in the ICCS in 2009 measuring the perceptions of school contexts, and the Love Relationship Scale for couples – were analyzed with the newly developed models and HCA. Results showed that clustered samples exhibited various

degrees of LPD across items, and fitting standard models disregarding LPD led to shrinkage scales. HCA on person residuals was applicable to detecting LPD, but its performance, especially with the paired samples, was unsatisfactory. The new models again demonstrate efficiency in quantifying the magnitudes of LPD over HCA.

6.2 Limitations and future research

This study adopted a simple method of adding a common parameter for persons within a cluster to account for the inter-person interaction on each item, i.e., in the proposed models, the inter-person interaction was treated as an all-channel network. Such imagination is reasonable in some cases (e.g., large-scale tests), but may not always be applicable to other scenarios. Other types of network patterns are possible (Ramos, 2012), and different levels of centralization can be found in different scenarios. For instance, in the studies of policy implementation within a community, a hierarchical structure is commonly found, in which a few grassroots officials possess an abundance of resources and authority to influence the behaviors of other residents (e.g., Li & O'Brien, 1999). Logically, it is more appropriate to consider the hierarchy within a cluster. In parenting studies, the effect of parenting is unidirectional so that the network of family members is more like a chain (Baumrind, 1971, 1991). With more variables describing person interactions, it is feasible to develop several delicate models that incorporate the most appropriate network patterns, according to each testing situation, for comparing the similarities and differences of parameter estimates under different models. In addition, the hierarchy in the GPD was not included in the proposed models. A multilevel structure also can be applied to LPD parameters to illustrate the effect of cluster-level covariates on LPD. How to deal with dual local dependence (i.e., coexistence of LID and LPD) in IRT is also a major problem for future studies.

As pointed out in Chapter 3, the proposed models include a large number of random variables to account for LPD (one random variable for each item), making marginal maximum likelihood estimation inapplicable. Bayesian methods with WinBUGS were, thus, applied in the study. Unfortunately, the maintenance of WinBUGS has been suspended since 2007. Therefore, the alternative OpenBUGS (Spiegelhalter, Thomas, Best, & Lunn, 2014) is recommended for future studies, as it also is an open-source package application developed by the MRC Biostatistics Unit in Cambridge for Bayesian analysis and is as efficient and reliable as WinBUGS over a wide range of applications. On the other hand, JAGS (Just Another Gibbs Sampler; Plummers, 2015) and NIMBLE (NIMBLE Development Team, 2016), which are both open-resource applications written in C++ for Bayesian analysis, also can be used. It is worthy to note that these packages were developed for general statistical models, i.e., they are not specialized for any statistical model, so the efficiency of calibration would be mediocre. Consequently, developing an optimized program to accelerate efficiency is essential for further studies.

The utility of HCA in accessing LPD was exemplified in this study, but two major limitations were raised in both the simulations and empirical examples. First, HCA on person residuals does not guarantee a high likelihood of recovering inputted clusters. In reality, a dataset can include numerous clusters exhibiting different levels of LPD. The reason why there was a mismatch between the explored clusters and true clusters in the examples is that the selected observations happened to exhibit similar patterns of LPD, such that the homogeneity of person residuals could not provide sufficient information for recovering true clusters. Secondly, many empirical situations restrict the applicability of cluster analysis. An apparent weakness is to conduct clusters with few observations, such as paired samples, then the conventional analysis becomes ineffective. In addition, occasionally, too many combinations of pre-matching clusters are available, making the use

of cluster analysis cumbersome. In sum, using HCA to assess LPD among person residuals is less efficient and provides barely satisfactory power compared with modeling LPD directly. One may try other conventional tools or develop a new statistic for detecting LPD and compare performance.

Similar to other simulation studies, the simulated conditions in this study were limited. One could include more comprehensive conditions (e.g., test lengths, non-normal distribution of the latent trait, and mixed format items) to investigate the performance of the new models in future studies. For example, varied combinations of clusters and the number of units are empirically possible. By following the research design in simulation study 2, a follow-up simulation using 250 couples answering 10 items was conducted to examine the applicability of the new models to minimum units (i.e., 2) within a cluster. The parameter recovery went fairly well, implying that LPD parameters can be estimated accurately, as long as the number of clusters is large. Herein, the minimum number of clusters is not recommended. As illustrated in the literature, the smaller the size, the more inaccurate the parameter estimates are. The guideline regarding the number of clusters when applying multilevel analyses could be a reference index. Conversely, researchers can fit the new models to their data and see whether the precision of parameter estimates is acceptable. Furthermore, applications of the new models with more empirical examples are needed to investigate the cause and influence of LPD in various situations.

Further model extensions are possible. For example, similar to modern IRT models, one can consider variant-item discrimination in the presented models or latent group membership at class/individual levels. Finally, the LPD modeling approach can be applied beyond the scope of IRT models. For example, compared with the latent trait approach, cognitive diagnostic models (CDMs), which are specialized to assess respondents' mastery of involved attributes, are gradually implemented in educational tests (de la Torre, 2011;

Haertel, 1989; Henson, Templin & Willse, 2009; Junker & Sijtsma, 2001; Templin & Henson, 2006; von Davier, 2008, 2013). Although distinct viewpoints toward measured proficiency are adopted in IRT and CDM, the analyzed dataset is the same. Therefore, the possibility of the existence of LPD should be carefully considered in CDMs. Researchers have incorporated multilevel modeling (e.g., Ayers, Rabe-Hesketh, & Nugent, 2013) or DIF (e.g., Li & Wang, 2015) into standard CDMs. Thus, one can incorporate these two components jointly into CDMs. For example, in the framework of the log-linear cognitive diagnosis model (LCDM; Henson, Templin, & Willse, 2009), the generalized model can be formulated as:

$$P_{ngi} = \frac{\exp\left(\lambda_{i,0} + \boldsymbol{\lambda}_i^{\mathrm{T}} h(\boldsymbol{\alpha}_{ng}, \mathbf{q}_i) + \xi_{ig}\right)}{1 + \exp\left(\lambda_{i,0} + \boldsymbol{\lambda}_i^{T} h(\boldsymbol{\alpha}_{ng}, \mathbf{q}_i) + \xi_{ig}\right)}, \tag{40}$$

in which $\boldsymbol{\alpha}_{ng}$ is the profile of person n within cluster g, including K latent attributes; $\lambda_{i,0}$ is the probability of success for test-takers who do not have any attributes measured by item i; $\boldsymbol{\lambda}_i^{\mathrm{T}}$ is a vector of weights for item i; $\mathbf{q}_i$ is a vector specifying the required attributes on item i; $h(\boldsymbol{\alpha}_{ng}, \mathbf{q}_i)$ is the combination of $\boldsymbol{\alpha}_{ng}$ and $\mathbf{q}$; and ξ_{ig} stands for the LPD among persons within cluster g on item i. The latent attribute α_{kng} can be assumed to follow a Bernoulli distribution, with a probability of π_{kng}, and π_{kng} is further modeled by logistic regression:

$$\pi_{kng} = \frac{\exp(\mathbf{b}'\mathbf{X}_{ng} - \delta_k)}{1 + \exp(\mathbf{b}'\mathbf{X}_{ng} - \delta_k)}, \tag{41}$$

in which $\mathbf{X}_{ng}$ is a vector, including individual and group-level covariates for person n within cluster g. Consequently, the GPD on attribute α_k can be quantified by including cluster-level covariates.

6.3 Conclusions

The proposed LPD modeling approach, as illustrated in Chapter 3, is technically the composition of multilevel modeling, a set of random variables for persons within a cluster

on all items, and a standard IRT model. Thus, it can be applied to numerous studies, including a person-clustering structure for examining how homogeneous the intended-to-be-measured latent trait(s) and response patterns among respondents within a cluster are. Compared with the standard multilevel models, the homogeneity of patterns from respondents within a cluster (i.e., LPD) is especially noticeable in the proposed models. The imagination and expectation of LPD depend on the contexts. On the one hand, LPD can be comprehended from the viewpoint of DIF. The magnitudes of LPD should be minimized in a fair measurement. If the magnitudes are too large to ignore, it is better to identify possible sources of DIF items for further revision or to remove them so that the test scores can be compared across clusters. On the other hand, LPD is anticipated in some situations. As demonstrated in Example 4, it is meaningful for couples that some of their thoughts and attitudes toward love were dependent on each other. Apparently, the active measure to explore LPD helps boost understanding of data.

This study illustrates that ignoring LPD by fitting standard models would lead to biased estimation and inflated GPD. The LPD modeling approach can be easily extended and applied to data with clustered samples; however, fitting a more complicated model arbitrarily comes with a price. It would increase the difficulty in interpreting test scores and decrease the generalizability of the research findings. When the magnitudes of LPD are found to be trivial for all items, standard multilevel models should suffice. Undoubtedly, it is much easier to explore the magnitudes of LPD than to explain its causes. The proposed methods simply come up with some quantitative evidence for LPD. The views of content experts are still needed to provide meaningful explanations as to why LPD occurs.

References

Adams, R. J., Wilson, M., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1-23. doi: 10.1177/0146621697211001

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561-573. doi: 10.1007/BF02293814

Angoff, W. H., & Ford, S. F. (1973). Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement, 10*, 95-106. doi: 10.1111/j.1745-3984.1973.tb00787.x

Ayers, E., Rabe-Hesketh, S., & Nugent, R. (2013). Incorporating student covariates in cognitive diagnosis models. *Journal of Classification, 30*, 195-224. doi: 10.1007/s00357-013-9130-y

Basturk, R. (2008). Applying the many-facet Rasch model to evaluate PowerPoint presentation performance in higher education. *Assessment & Evaluation in Higher Education, 33*, 431-444. doi: 10.1080/02602930701562775

Bassili, J. N., & Scott, S. B. (1996). Response latency as a signal to question problems in survey research. *Public Opinion Quarterly, 60*, 390-399. doi:10.1086/297760

Baumrind, D. (1971). Current patterns of parental authority. *Developmental Psychology, 4*, 1-103. doi: 10.1037/h0030372

Baumrind, D. (1991). The influence of parenting style on adolescent competence and substance use. The Journal of Early Adolescence, 11, 56-95. doi: 10.1177/0272431691111004

Belov, D. I. (2011). Detection of answer copying based on the structure of a high-stakes test. *Applied Psychological Measurement, 35*, 495-517. doi: 10.1177/0146621611420705

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's

ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley.

Bock, R. D. (Ed.). (1989). *Multilevel analysis of educational data*. San Diego, CA: Academic Press.

Briggs, D. C., & Wilson, M. (2003). An introduction to multidimensional measurement using Rasch models. *Journal of Applied Measurement, 4*, 87-100.

Brophy, J. (2006). History of Research on Classroom Management. In C.M. Evertson & C.S. Weinstein (Eds.), *Handbook of Classroom Management: Research, Practice, and Contemporary Issues* (pp. 17–43). USA: Lawrence Erlbaum Associates.

Brown, T. A. (2015). *Confirmatory Factor Analysis for Applied Research*. New York, NY: Guilford Publications.

Burstein, L. (1980). The analysis of multilevel data in educational research and evaluation. *Review of Research in Education, 8*, 158-233. doi: 10.2307/1167125

Chen, C.-T., & Wang, W.-C. (2007). Effects of ignoring item interaction on item parameter estimation and detection of interacting items. *Applied Psychological Measurement, 31*, 388-411. doi: 10.1177/0146621606297309

Chessa, A. G., & Holleman, B. C. (2007). Answering attitudinal questions: Modelling the response process underlying contrastive questions. *Applied Cognitive Psychology, 21*, 203-225. doi:10.1002/acp.1337

Cho, S.-J., & Cohen, A. S. (2010). A multilevel mixture IRT model with an application to DIF. *Journal of Educational and Behavioral Statistics, 35*, 336-370. doi: 10.3102/1076998609353111

Cho, S.-J. Suh, Y., & Lee, W.-Y. (2016). After differential item functioning is detected: IRT item calibration and scoring in the presence of DIF. *Applied Psychological Measurement, 40*, 573-591. doi: 10.1177/0146621616664304

Cizek, G. J. (1999). *Cheating on tests: How to do it, detect it, and prevent it*. Mahwah, NJ: Lawrence Erlbaum.

Cohen, A. S., Kim, S.-H., & Wollack, J. A. (1996). An investigation of the likelihood ratio test for detection of differential item functioning. *Applied Psychological Measurement, 20*, 15-26. doi: 10.1177/014662169602000102

Congdon, P. J., & MeQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement, 37*, 163-178. doi: 10.1111/j.1745-3984.2000.tb01081.x

Cristante, F., & Robusto, E. (1999). Assessing dependence among subjects' responses. *Mathematical Social Sciences, 38*, 259-274. doi: 10.1016/S0165-4896(99)00020-7

de Jong, M. G., & Steenkamp, J.-B. E. M. (2010). Finite mixture multilevel multidimensional ordinal IRT models for large scale cross-cultural research. *Psychometrika, 75*, 3-32. doi: 10.1007/s11336-009-9134-z

de Jong, M. G., Steenkamp, J.-B. E. M., & Fox, J.-P. (2007). Relaxing measurement invariance in cross-national consumer research using a hierarchical IRT model. *Journal of Consumer Research, 34*, 260-278. doi: 10.1086/518532

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*, 179-199. doi: 10.1007/s11336-011-9207-7

de la Torre, J., & Hong, Y. (2010). Parameter estimation with small sample size a higher-order IRT model approach. *Applied Psychological Measurement, 34*, 267-285. doi: 10.1177/0146621608329501

de la Torre, J., & Song, H. (2009). Simultaneously estimation of overall and domain abilities: A higher-order IRT model approach. *Applied Psychological Measurement, 33*, 620-639. doi: 10.1177/0146621608326423

Deal, J. E. (1995). Utilizing data from multiple family members: A within-family approach.

Journal of Marriage and Family, 57, 1109-1121. doi: 10.2307/353426

DeVellis, R. F. (2005). *Scale development: Theory and applications* (2nd ed.). Thousand Oaks, CA: Sage.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23*, 355-368. doi: 10.1111/j.1745-3984.1986.tb00255.x

Du, Y., Wright, B. D., & Brown, W. L. (1996). *Differential facet functioning detection in direct writing assessment*. Paper presented at the Annual Meeting of the American Educational Research Association, New York, USA.

Eberbach, C., & Crowley, K. (2009). From everyday to scientific observation: How children learn to observe the biologist's world. *Review of Educational Research, 79*, 39-68. doi: 10.3102/0034654308325899

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly, 2*, 197-221. doi: 10.1207/s15434311laq0203_2

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25*, 155-185. doi: 10.1177/0265532207086780

Engelhard, G. J. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement, 31*, 93-112. doi: 10.1111/j.1745-3984.1994.tb00436.x

Engelhard, G. J. (1996). Evaluating rater accuracy in performance assessments. *Journal of Educational Measurement, 33*, 56-70. doi: 10.1111/j.1745-3984.1996.tb00479.x

Engelhard, G. J. (2008). Differential rater functioning. *Rasch Measurement Transactions, 21*, 1124. Retrieved from http://www.rasch.org/rmt/rmt213f.htm

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, N. J.: Erlbaum Publishers.

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement, 29*, 278-295. doi: 10.1177/0146621605275728

Finch, W. H. (2016). Detection of differential item functioning for more than two groups: A monte carlo comparison of methods. *Applied Measurement in Education, 29*, 30-45. doi: 10.1080/08957347.2015.1102916

Fox, J.-P. (2005). Multilevel IRT using dichotomous and polytomous response data. *British Journal of Mathematical and Statistical Psychology, 58*, 145-172. doi: 10.1348/000711005x38951

Fox, J.-P., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using gibbs sampling. *Psychometrika, 66*, 271-288. doi: 10.1007/bf02294839

French, B. F., & Finch, W. H. (2010). Hierarchical logistic regression: Accounting for multilevel data in DIF detection. *Journal of Educational Measurement, 47*, 299-317. doi: 10.1111/j.1745-3984.2010.00115.x

French, B. F., & Finch, W. H. (2013). Extensions of Mantel–Haenszel for multilevel DIF detection. *Educational and Psychological Measurement, 73*, 648-671. doi: 10.1177/0013164412472341

Ganong, L. H. (2003). Selecting family measurements. *Journal of Family Nursing, 9*, 184-206. doi: 10.1177/1074840703009002005

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistical Sinica, 6*, 733-807.

Glas, C. A. W., & Pimentel, J. L. (2008). Modeling nonignorable missing data in speeded tests. *Educational and Psychological Measurement, 68*, 907-922. doi:

10.1177/0013164408315262

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*, 301-321. doi: 10.2307/1434756

Haarr, R. N. (2003). Impact of Community Policing Training and Program Implementation on Police Personnel in Arizona, 1995-1998: Inter-university Consortium for Political and Social Research (ICPSR) [distributor].

Harris, K. M., & Udry, J. R. (2016). National Longitudinal Study of Adolescent to Adult Health (Add Health), 1994-2008 [Public Use]: Inter-university Consortium for Political and Social Research (ICPSR) [distributor].

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74*, 191-210. doi: 10.1007/s11336-008-9089-5

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test Validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Holland, P. W., & Wainer, H. (1993). Differential item functioning. Hillsdale, NJ: Erlbaum.

Holman, R., & Glas, C. A. W. (2005). Modelling non-ignorable missing-data mechanisms with item response theory models. *British Journal of Mathematical and Statistical Psychology, 58*, 1-17. doi: 10.1111/j.2044-8317.2005.tb00312.x

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2 ed.). New York, NY: Routledge.

Hox, J. J., & Roberts, J. K. (2011). Multilevel analysis: Where we were and where we are. In J. J. Hox & J. K. Roberts (Eds.), *Handbook of Advanced Multilevel Analysis* (pp. 3-11). New York, NY: Routledge.

Huang, H.-Y., & Wang, W.-C. (2013). Higher-order testlet response models for hierarchical latent traits and testlet-based items. *Educational and Psychological Measurement, 73*, 491-511. doi: 10.1177/0013164412454431

Huang, H.-Y., Wang, W.-C, Chen, P.-H, & Su, C.-M. (2013). Higher-order item response models for hierarchical latent traits. *Applied Psychological Measurement, 37*, 619-637. doi: 10.1177/0146621613488819

Hung, L.-F. (2011). Formulation and application of the hierarchical generalized random-situation random-weight MIRID. Multivariate Behavioral Research, 46, 643-668. doi: 10.1080/00273171.2011.589274

Jager, J., Bornstein, M. H., Putnick, D. L., & Hendricks, C. (2012). Family members' unique perspectives of the family: Examining their scope, size, and relations to individual adjustment. *Journal of Family Psychology, 26*, 400-410. doi: 10.1037/a0028330

Jiao, H., Kamata, A., Wang, S., & Jin, Y. (2012). A multilevel testlet model for dual local dependence. *Journal of Educational Measurement, 49*, 82-100. doi: 10.1111/j.1745-3984.2011.00161.x

Jin, K-.Y., Chen, H.-F., & Wang, W.-C. (2015). Assessing differential item functioning in multiple grouping variables with factorial logistic regression. In R. E. Millsap, D. M. Bolt, L. A. van der Ark & W.-C. Wang (Eds.), *Quantitative Psychology Research: The 78th Annual Meeting of the Psychometric Society* (Vol. 89, pp. 243-259). Cham: Springer International Publishing.

Jin, K.-Y., & Wang, W.-C. (2016). Item response theory models for person dependence in paired samples. In L. A. van der Ark, D. M. Bolt, W.-C. Wang, J. A. Douglas & M. Wiberg (Eds.), *Quantitative Psychology Research: The 80th Annual Meeting of the Psychometric Society* (pp. 105-121). Cham: Springer International Publishing.

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from Web-based

personality inventories. *Journal of Research in Personality, 39*, 103-129. doi: 10.1016/j.jrp.2004.09.009

Johnson, M. S. (2007). Marginal maximum likelihood rstimation of item response models in R. *Journal of Statistical Software, 20*(10), 1-24.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*, 258-272. doi: 10.1177/01466210122032064

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement, 38*, 79-93. doi: 10.1111/j.1745-3984.2001.tb01117.x

Katama, A., & Vaughn, B. K. (2011). Multilevel IRT Modeling. In J. J. Hox & J. K. Roberts (Eds.), *Handbook of Advanced Multilevel Analysis* (pp. 41-58). New York, NY: Routledge.

Kamoen, N., Holleman, B., Mak, P., Sanders, T., & van den Bergh, H. (2011). Agree or disagree? Cognitive processes in answering contrastive survey questions. *Discourse Processes, 48*, 355-385. doi:10.1080/0163853X.2011.578910

Kieruj, N. D., & Moors, G. (2013). Response style behavior: Question format dependent or personal style. *Quality & Quantity, 47*, 193-211. doi:10.1007/s11135-011-9511-4

Kim, S., Cohen, A., & Park, T. (1995). Detection of differential item functioning in multiple groups. *Journal of Educational Measurement, 32*, 261-276. doi: 10.1111/j.1745-3984.1995.tb00466.x

Kish, L. (1965). *Survey sampling*. New York: John Wiley.

Langford, I. H., Leyland, A. H., Rasbash, J., & Goldstein, H. (1999). Multilevel modelling of the geographical distributions of diseases. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 48*, 253-268.

Li, L., & O'brien, K. J. (1999). Selective policy implementation in rural China.

Comparative Politics, 31, 167-186.

Li, T., Jiao, H., & Macready, G. B. (2016). Different approaches to covariate inclusion in the mixture Rasch model. *Educational and Psychological Measurement, 76*, 848-872. doi: 10.1177/0013164415610380

Li, X., & Wang, W.-C. (2015). Assessment of differential item functioning under cognitive diagnosis models: The DINA model example. *Journal of Educational Measurement, 52*, 28-54. doi: 10.1111/jedm.12061

Linacre, J. M. (2016). Winsteps® (Version 3.92.0) [Computer Software]. Beaverton, Oregon: Winsteps.com. Retrieved January 1, 2016. Available from http://www.winsteps.com/

Lord, F. M. (1980). *Application of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing, 10*, 325-337.

Maier, K. S. (2001). A Rasch hierarchical measurement model. *Journal of Educational and Behavioral Statistics, 26*, 307-330. doi: 10.3102/10769986026003307

Magis, D., Béland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods, 42*, 847-862. doi: 10.3758/brm.42.3.847

Magis, D., Raîche, G., Béland, S., & Gérard, P. (2011). A generalized logistic regression procedure to detect differential item functioning among multiple groups. *International Journal of Testing, 11*, 365-386. doi:10.1080/15305058.2011.602810

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*,

719-748. doi: 10.1093/jnci/22.4.719

Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174. doi: 10.1007/BF02296272

Matin, L., & Adkins, D. (1954). A second-order factor analysis of reasoning abilities. *Psychometrika, 19*, 71-78. doi: 10.1007/BF02288995

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods, 17*, 437-455. doi: 10.1037/a0028085

Menezes, N. M., Georgiades, K., & Boyle, M. H. (2011). The influence of immigrant status and concentration on psychiatric disorder in Canada: A multi-level analysis. *Psychological Medicine, 41*, 2221-2231. doi:10.1017/S0033291711000213

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*, 177-196. doi: 10.1007/bf02294457

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176. doi: 10.1177/014662169201600206

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4*, 386-422.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement, 5*, 189-227.

National Center for Family and Marriage Research, Diamond, L., & Hicks, A. (2010). Familial responses to financial instability, "It's all your fault": Predictors and implications of blame in couples under economic strain, 2009 [United States]. ICPSR26544-v1. Ann Arbor, MI: Inter-university Consortium for Political and

Social Research [distributor], 2010-05-20. doi: 10.3886/ICPSR26544.v1

Nichols, D. S., Greene, R. L., & Schmolck, P. (1989). Criteria for assessing inconsistent patterns of item endorsement on the MMPI: rationale, development, and empirical trials. *Journal of Clinical Psychology, 45*, 239-250. doi: 10.1002/1097-4679(198903)45:2<239::AID-JCLP2270450210>3.0.CO;2-1

NIMBLE Development Team. (2016). NIMBLE: An R Package for Programming with BUGS models (Version 0.6-3) [Computer software]. Retrieved from http://r-nimble.org.

Olsbjerg, M., & Christensen, K.B. (2014). Modeling local dependence in longitudinal IRT models. Behavior Research Methods, 47, 1413-1424. doi: 10.3758/s13428-014-0553-0

Opdenakker, M.-C., Van Damme, J., De Fraine, D. F., Van Landeghem, G., & Onghena, P. (2002). The effect of schools and classes on mathematics achievement. *School Effectiveness and School Improvement, 13*, 399-427. doi: 10.1076/sesi.13.4.399.10283

Organization for Economic Cooperation and Development. (2014). *PISA 2012 technical report*. Paris, France: Organization for Economic Cooperation and Development. Retrieved from http://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf

Penfield, R. D. (2001). Assessing differential item functioning among multiple groups: A comparison of three MantelHaenszel procedures. *Applied Measurement in Education, 14*, 235-259. doi:10.1207/S15324818AME1403_3

Plummer, M. (2015). JAGS version 4.0.0 user manual. Technical report. Retrieved from http://www.uvm.edu/~bbeckage/Teaching/DataAnalysis/Manuals/manual.jags.pdf

R Core Team (2016). *R: A language and environment for statistical computing*. Vienna,

Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Ramos, P. P. (2012). *Network models for organizations*. New York, NY: Palgrave Macmillan.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen: Institute of Education Research.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53*, 495-502. doi: 10.1007/bf02294403

Raju, N. S. (1990). Determining the significance of estimated signed and Uunsigned areas between two item response functions. *Applied Psychological Measurement, 14*, 197-207. doi: 10.1177/014662169001400208

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17*, 105-116. doi: 10.1177/014662169301700201

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

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing, 25*, 465-493. doi: 10.1177/0265532208094273

Schultz, W., Ainley, J., & Fraillon, J. (Eds.). (2011). *ICCS 2009 Technical Report*. Amsterdam, The Netherlands: International Association for the Evaluation of Educational Achievement (IEA).

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58*, 159-194. doi: 10.1007/bf02294572

Sinharay, S. (2005). Assessing fit of unidimensional item response theory models using a

bayesian approach. *Journal of Educational Measurement, 42*, 375-394. doi: 10.1111/j.1745-3984.2005.00021.x

Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement, 30*, 298-321. doi: 10.1177/0146621605285517

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28*, 237-247. doi: 10.1111/j.1745-3984.1991.tb00356.x

Sheng, Y., & Wikle, C. K. (2008). Bayesian multidimensional IRT models with a hierarchical structure. *Educational and Psychological Measurement, 68*, 413-430. doi: 10.1177/0013164407308512

Spiegelhalter D., Thomas A., Best N., & Lunn D. J. (2014). *OpenBUGS user manual: Version 3.2.3*. Retrieved from http://www.mrc-bsu.cam.ac.uk/bugs

Steenkamp, J.-B. E. M, & Baumgartner, H. (1998). Assessing Measurement Invariance in Cross-National Consumer Research. *Journal of Consumer Research, 25*, 78-107. doi:10.1086/209528

Stephen, W. R., & Anthony, S. B. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods* (2 ed.). Newbury Park, CA: SAGE Publications.

Sternberg, R. J. (1986). A triangular theory of love. *Psychological Review, 92*, 119-135. doi: 10.1037/0033-295X.93.2.119

Sternberg, R. J. (1987). Liking versus loving: A comparative evaluation of theories. *Psychological Bulletin, 102*, 331-345. doi: 10.1037/0033-2909.102.3.331

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*, 361-370. doi: 10.1111/j.1745-3984.1990.tb00754.x

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*, 287-305. doi: 10.1037/1082-989X.11.3.287

Tuerlinckx, F., & De Boeck, P. (2001). The effect of ignoring item interactions on the estimated discrimination parameters in item response theory. *Psycholotical Methods, 6*, 181-195. doi: 10.1037/1082-989X.6.2.181

Tuerlinckx, F., Rijmen, F., Molenberghs, G., Verbeke, G., Briggs, D. C., Van der Noortgate, W., . . . De Boeck, P. (2004). Estimation and software. In P. De Boeck & M. Wilson (Eds.), Explanatory item response models: A generalized linear and nonlinear approach. New York: Springer.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61*, 287-307. doi: 10.1348/000711007x193957

von Davier, M. (2014). The DINA model as a constrained general diagnostic model: Two variants of a model equivalency. *British Journal of Mathematical and Statistical Psychology, 67*, 49-71. doi: 10.1111/bmsp.12003

Wainer, H., Bradlow, E. T., & Du, Z. (2000). Testlet response theory: An analog for the 3PL model useful in testlet-based adaptive testing. In W. J. van der Linden & G. A. W. Glas (Eds.), *Computerized Adaptive Testing: Theory and Practice* (pp. 245-269). Dordrecht: Springer Netherlands.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications.* New York: Cambridge University Press.

Wang, C.-F. (2000). *The matching of attachment styles and its influence to love relationship and relationship adjustment (E88016)*. [Data file]. Available from Survey Research Data Archive, Academia Sinica. doi:10.6141/TW-SRDA-E88016-1

Wang, W.-C. (2000).The simultaneous factorial analysis of differential item functioning. *Methods of Psychological Research, 5*, 56-76.

Wang, W.-C., Chen, H.-F., & Jin, K.-Y. (2015). Item response theory models for wording effects in mixed-format scales. *Educational and Psychological Measurement, 75*, 157-178. doi: 10.1177/0013164414528209

Wang, W.-C., & Jin, K.-Y. (2016). Analysis of testlet data. In Q. Zhang, *Pacific Rim Objective Measurement Symposium (PROMS) 2015 Conference Proceedings* (pp. 199-241): Springer Singapore.

Wang, W.-C., & Shih, C.-L. (2010). MIMIC methods for assessing differential item functioning in polytomous items. *Applied Psychological Measurement, 34*, 166-180. doi: 10.1177/0146621609355279

Wang, W.-C., Shih, C.-L., & Yang, C.-C. (2009). The MIMIC method with scale purification for detecting differential item functioning. *Educational and Psychological Measurement, 69*, 713-731. doi: 10.1177/0013164409332228

Wang, W.-C., Su, C.-M., & Qiu, X.-L. (2014). Item response models for local dependence among multiple ratings. *Journal of Educational Measurement, 53*, 260-280. doi: 10.1111/jedm.12045

Wang, W.-C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement, 29*, 126-149. doi: 10.1177/0146621604271053

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association, 58(301)*, 236-244.

Wilson, M. (2004). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates.

Wollack, J. A., Cohen, A. S., & Eckerly, C. A. (2015). Detecting test tampering using item response theory. *Educational and Psychological Measurement, 75*, 931-935. doi:

10.1177/0013164414568716

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*, 125-145. doi: 10.1177/014662168400800201

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187-213. doi: 10.1111/j.1745-3984.1993.tb00423.x

Appendix A: WinBUGS Codes for the MPCM-CS of Simulation Studies 1 and 2

```
# N is the number of examinees;
# T is the number of items;
# r is the data matrix with N rows and T columns;
# idx is the cluster index;
# lv1.p is the person-level predictor (i.e., -1 or 1), and lv1.pc is the predicted coefficient;
# lv2.p is the cluster-level predictor (i.e., -1 or 1), and lv2.pc is the predicted coefficient;
# theta is the person ability;
# delta are the item thresholds;
# and xi is the LPD parameter.

model {
    for (i in 1:N) {
        lv1.e[i] ~ dnorm(0, tau.lv1)
        theta[i] <- (lv2.pc[lv2.p[i]] + lv2.e[idx[i]])+ lv1.pc[lv1.p[i]] + lv1.e[i]
        for (j in 1:T) {
            Q[i,j,1] <- 1
            Q[i,j,2] <- exp(theta[i] - delta[j,1] + xi[idx[i],j])
            Q[i,j,3] <- exp(2*theta[i] - delta[j,1] - delta[j,2] + 2*xi[idx[i],j])
            Q[i,j,4] <- exp(3*theta[i] - delta[j,1] - delta[j,2] - delta[j,3] + 3*xi[idx[i],j])
            denom[i,j] <- sum(Q[i,j,])
            PP[i,j,1] <- Q[i,j,1]/denom[i,j]
            PP[i,j,2] <- Q[i,j,2]/denom[i,j]
            PP[i,j,3] <- Q[i,j,3]/denom[i,j]
            PP[i,j,4] <- Q[i,j,4]/denom[i,j]
            r[i,j] ~ dcat(PP[i,j,])
        }
    }
# Priors
    tau.lv1 ~ dgamma(0.1, 0.1)
    sigma.lv1 <- 1/tau.lv1
    tau.lv2 ~ dgamma(0.1, 0.1)
    sigma.lv2 <- 1/tau.lv2
    lv1.pc[1] ~ dnorm(0, 0.1)
    lv1.pc[2] <- -lv1.pc[1]
    lv2.pc[1] ~ dnorm(0, 0.1)
    lv2.pc[2] <- -lv2.pc[1]
    for (i in 1:100) {     # there are 100 clusters
```

```
            lv2.e[i] ~ dnorm(0, tau.lv2)
            for (j in 1:T) {
                xi[i,j] ~ dnorm(0, tau.xi[j])
            }
        }
        for (j in 1:T) {
            delta[j,1] ~ dnorm(0, 0.1)
            delta[j,2] ~ dnorm(0, 0.1)
            delta[j,3] ~ dnorm(0, 0.1)
            tau.xi[j] ~ dgamma(0.1, 0.1)
            sigma.xi[j] <- 1/tau.xi[j]
        }
    }
```

Appendix B: WinBUGS Codes for the MDMPCM-CS of Simulation Study 3

```
# N is the number of examinees;
# T is the number of items;
# r is the data matrix with N rows and T columns;
# idx is the cluster index;
# lv1.p is the person-level predictor (i.e., -1 or 1), and lv1.pc is the predicted coefficient;
# theta is the person ability;
# delta are the item thresholds;
# and xi is the LPD parameter.

model {
    for (i in 1:N) {
        theta.i[i,1:3] ~ dmnorm(mu.i[1:3], I_cov.i[1:3, 1:3])
        for (d in 1:3) {
            theta[i,d] <- theta.i[i,d] + lv1.pc[d,lv1.p[i]] + theta.s[idx[i],d]
            for (j in (5*(d-1)+1): (5*d)) {
                Q[i,j,1] <- 1
                Q[i,j,2] <- exp(theta[i,d] - delta[j,1] + xi[idx[i],j])
                Q[i,j,3] <- exp(2*theta[i,d] - delta[j,1] - delta[j,2] + 2*xi[idx[i],j])
                Q[i,j,4] <- exp(3*theta[i,d] - delta[j,1] - delta[j,2] - delta[j,3] +
3*xi[idx[i],j])
                denom[i,j] <- sum(Q[i,j,])
                PP[i,j,1] <- Q[i,j,1]/denom[i,j]
                PP[i,j,2] <- Q[i,j,2]/denom[i,j]
                PP[i,j,3] <- Q[i,j,3]/denom[i,j]
                PP[i,j,4] <- Q[i,j,4]/denom[i,j]
                r[i,j] ~ dcat(PP[i,j,])
            }
        }
    }
# Priors
    mu.i[1] <- 0
    mu.i[2] <- 0
    mu.i[3] <- 0
    I_cov.i[1:3, 1:3] ~ dwish(alpha[1:3, 1:3], 3)
    covm.i[1:3, 1:3] <- inverse(I_cov.i[1:3, 1:3])
    rho.i[1,2] <- covm.i[1,2]/sqrt(covm.i[1,1]*covm.i[2,2])
    rho.i[1,3] <- covm.i[1,3]/sqrt(covm.i[1,1]*covm.i[3,3])
```

```
        rho.i[2,3] <- covm.i[2,3]/sqrt(covm.i[2,2]*covm.i[3,3])
        mu.s[1] <- 0
        mu.s[2] <- 0
        mu.s[3] <- 0
        I_cov.s[1:3, 1:3] ~ dwish(alpha[1:3, 1:3], 3)
        covm.s[1:3, 1:3] <- inverse(I_cov.s[1:3, 1:3])
        rho.s[1,2] <- covm.s[1,2]/sqrt(covm.s[1,1]*covm.s[2,2])
        rho.s[1,3] <- covm.s[1,3]/sqrt(covm.s[1,1]*covm.s[3,3])
        rho.s[2,3] <- covm.s[2,3]/sqrt(covm.s[2,2]*covm.s[3,3])
        lv1.pc[1,1] ~ dnorm(0, 0.1)
        lv1.pc[1,2] <- -lv1.pc[1,1]
        lv1.pc[2,1] ~ dnorm(0, 0.1)
        lv1.pc[2,2] <- -lv1.pc[2,1]
        lv1.pc[3,1] ~ dnorm(0, 0.1)
        lv1.pc[3,2] <- -lv1.pc[3,1]
        for (i in 1:200) {     # there are 200 clusters
            theta.s[i,1:3] ~ dmnorm(mu.s[1:3], I_cov.s[1:3, 1:3])
            for (j in 1:T) {
                xi[i,j] ~ dnorm(0, tau.xi[j])
            }
        }
        for (j in 1:T) {
            delta[j,1] ~ dnorm(0, 0.1)
            delta[j,2] ~ dnorm(0, 0.1)
            delta[j,3] ~ dnorm(0, 0.1)
            tau.xi[j] ~ dgamma(0.1, 0.1)
            sigma.xi[j] <- 1/tau.xi[j]
        }
    }
```

Appendix C: WinBUGS Codes for the HOPCM-CS of Simulation Study 4

```
# N is the number of examinees;
# T is the number of items;
# r is the data matrix with N rows and T columns;
# idx is the cluster index;
# lambda is the regression weight of the 2nd-order latent trait on the 1st-order latent trait;
# nu is the regression error;
# lv1.p is the person-level predictor (i.e., -1 or 1), and lv1.pc is the predicted coefficient;
# lv2.p is the cluster-level predictor (i.e., -1 or 1), and lv2.pc is the predicted coefficient;
# theta is the person ability;
# delta are the item thresholds;
# and xi is the LPD parameter.

model {
    for (i in 1:N) {
        lv1.e[i] ~ dnorm(0, tau.lv1)
        theta[i,5] <- (lv2.pc[lv2.p[i]] + lv2.e[idx[i]])+ lv1.pc[lv1.p[i]] + lv1.e[i]

        for (d in 1:4) {
            nu[i,d] ~ dnorm(0, inv.nu[d])
            theta[i,d] <- lambda[d]*theta[i,5] + nu[i,d]
            for (j in (5*(d-1)+1): (5*d)) {
                Q[i,j,1] <- 1
                Q[i,j,2] <- exp(theta[i,d] - delta[j,1] + xi[idx[i],j])
                Q[i,j,3] <- exp(2*theta[i,d] - delta[j,1] - delta[j,2] + 2*xi[idx[i],j])
                Q[i,j,4] <- exp(3*theta[i,d] - delta[j,1] - delta[j,2] - delta[j,3] +
3*xi[idx[i],j])
                denom[i,j] <- sum(Q[i,j,])
                PP[i,j,1] <- Q[i,j,1]/denom[i,j]
                PP[i,j,2] <- Q[i,j,2]/denom[i,j]
                PP[i,j,3] <- Q[i,j,3]/denom[i,j]
                PP[i,j,4] <- Q[i,j,4]/denom[i,j]
                r[i,j] ~ dcat(PP[i,j,])
            }
        }
    }
# Priors
    lambda[1] <- 1
```

```
        lambda[2] ~ dlnorm(0, 0.1)
        lambda[3] ~ dlnorm(0, 0.1)
        lambda[4] ~ dlnorm(0, 0.1)
        inv.nu[1] ~ dgamma(0.1, 0.1)
        inv.nu[2] ~ dgamma(0.1, 0.1)
        inv.nu[3] ~ dgamma(0.1, 0.1)
        inv.nu[4] ~ dgamma(0.1, 0.1)
        sigma.nu[1] <- 1/inv.nu[1]
        sigma.nu[2] <- 1/inv.nu[2]
        sigma.nu[3] <- 1/inv.nu[3]
        sigma.nu[4] <- 1/inv.nu[4]
        tau.lv1 ~ dgamma(0.1, 0.1)
        sigma.lv1 <- 1/tau.lv1
        tau.lv2 ~ dgamma(0.1, 0.1)
        sigma.lv2 <- 1/tau.lv2
        lv1.pc[1] ~ dnorm(0, 0.1)
        lv1.pc[2] <- -lv1.pc[1]
        lv2.pc[1] ~ dnorm(0, 0.1)
        lv2.pc[2] <- -lv2.pc[1]
        for (i in 1:200) {     # there are 200 clusters
            lv2.e[i] ~ dnorm(0, tau.lv2)
            for (j in 1:T) {
                xi[i,j] ~ dnorm(0, tau.xi[j])
            }
        }
        for (j in 1:T) {
            delta[j,1] ~ dnorm(0, 0.1)
            delta[j,2] ~ dnorm(0, 0.1)
            delta[j,3] ~ dnorm(0, 0.1)
            tau.xi[j] ~ dgamma(0.1, 0.1)
            sigma.xi[j] <- 1/tau.xi[j]
        }
    }
```

Appendix D: WinBUGS Codes for the MFRM-CS of Simulation Study 5

```
# N is the number of examinees;
# T is the number of items;
# R is the number of raters;
# r is the data matrix with N × R rows and T + 3 columns, and the first three columns denote
person idex, cluster index, and rater index, respectively;
# id is the data matrix with N rows and 4 clomuns including person id, person-level
predictor (i.e., -1 or 1), group-level predictor (i.e., -1 or 1), and cluster index;
# lv1.pc is the predicted coefficient;
# lv2.pc is the predicted coefficient;
# theta is the person ability;
# delta are the item thresholds;
# iota is the rater severity;
# and xi is the LPD parameter.

model {
    for (i in 1:4000) {   # there are totally 4000 responses
        for (j in 1:T) {
            Q[r[i,1],j,r[i,3],1] <- 1     # Q[id, item, rater, category]
            Q[r[i,1],j,r[i,3],2] <- exp(theta[r[i,1]] - delta[j,1] - iota[r[i,3]] +
xi[r[i,2],r[i,3]])
            Q[r[i,1],j,r[i,3],3] <- exp(2*theta[r[i,1]] - delta[j,1] - delta[j,2] – 2*iota[r[i,3]]
+ 2*xi[r[i,2],r[i,3]])
            Q[r[i,1],j,r[i,3],4] <- exp(3*theta[r[i,1]] - delta[j,1] - delta[j,2] – delta[j,3] –
3*iota[r[i,3]] + 3*xi[r[i,2],r[i,3]])
            denom[r[i,1],j,r[i,3]] <- sum(Q[r[i,1],j,r[i,3],])
            PP[r[i,1],j,r[i,3],1] <- Q[r[i,1],j,r[i,3],1]/denom[r[i,1],j,r[i,3]]
            PP[r[i,1],j,r[i,3],2] <- Q[r[i,1],j,r[i,3],2]/denom[r[i,1],j,r[i,3]]
            PP[r[i,1],j,r[i,3],3] <- Q[r[i,1],j,r[i,3],3]/denom[r[i,1],j,r[i,3]]
            PP[r[i,1],j,r[i,3],4] <- Q[r[i,1],j,r[i,3],4]/denom[r[i,1],j,r[i,3]]
            r[i,j+3] ~ dcat(PP[r[i,1],j,r[i,3],])
        }
    }
    # Priors
    for (i in 1:N) {
        lv1.e[i] ~ dnorm(0, tau.lv1)
        theta[i] <- lv1.pc[id[i,2]] + (lv2.pc[id[i,3]] + lv2.e[id[i,4]]) + lv1.e[i]
    }
```

```
        tau.lv1 ~ dgamma(0.1, 0.1)
        sigma.lv1 <- 1/tau.lv1
        tau.lv2 ~ dgamma(0.1, 0.1)
        sigma.lv2 <- 1/tau.lv2
        lv1.pc[1] ~ dnorm(0, 0.1)
        lv1.pc[2] <- -lv1.pc[1]
        lv2.pc[1] ~ dnorm(0, 0.1)
        lv2.pc[2] <- -lv2.pc[1]
        for (i in 1:200) {     # there are 200 clusters
              lv2.e[i] ~ dnorm(0, tau.lv2)
              for (j in 1:R) {
                    xi[i,j] ~ dnorm(0, tau.xi[j])
              }
        }
        for (j in 1:T) {
              delta[j,1] ~ dnorm(0, 0.1)
              delta[j,2] ~ dnorm(0, 0.1)
              delta[j,3] ~ dnorm(0, 0.1)
              delta[j,4] ~ dnorm(0, 0.1)
        }
        tau.xi[1] ~ dgamma(0.1, 0.1)
        sigma.xi[1] <- 1/tau.xi[1]
        iota[1] <- 0
        for (j in 2:R) {
              iota[j] ~ dnorm(0, 0.1)
              tau.xi[j] ~ dgamma(0.1, 0.1)
              sigma.xi[j] <- 1/tau.xi[j]
        }
}
```

Appendix E: Item Descriptions in the Subscales Reflecting Students' Perceptions of the School Context in the ICCS 2009

Domain	Item	Description
Open schools	IS2G16B	Teachers encourage students to make up their own minds
	IS2G16C	Teachers encourage students to express their opinions
	IS2G16D	Students bring up current political events for discussion in class
	IS2G16E	Students express opinions in class even when their opinions are different from most of the other students
	IS2G16F	Teachers encourage students to discuss the issues with people having different opinions
	IS2G16G	Teachers present several sides of the issues when explaining them in class
Student influence	IS2G17A	The way classes are taught
	IS2G17B	What is taught in classes
	IS2G17C	Teaching and learning materials
	IS2G17D	The timetable
	IS2G17E	Classroom rules
	IS2G17F	School rules
Student-teacher relations	IS2G18A	Most of my teachers treat me fairly
	IS2G18B	Students get along well with most teachers
	IS2G18C	Most teachers are interested in students' wellbeing
	IS2G18E	Most of my teachers really listen to what I have to say
	IS2G18F	If I need extra help, I will receive it from my teachers

Appendix F: Item Descriptions in the Love Relationship Scale (Chinese version)

Domain	No.	Description
Passion	2	我發現自己一天之中常常想到他（她）。
	4	只要見到他（她），我就感到興奮快樂。
	5	我覺得他（她）的身體很有吸引力。
	9	我希望他（她）覺得我很有吸引力。
	11	跟他（她）在一起我感覺很浪漫興奮。
	19	即使他（她）不在身邊，我仍會常常想到他（她） 。
	23	只要注視著他（她），我就感到很快樂滿足。
	24	他（她）的一舉一動（一顰一笑）佔據了我的心思。
Intimacy	1	我與他（她）的關係是溫暖愉快的。
	3	當失意難過時，我能從他（她）得到適當的情緒支持。
	6	當他（她）失意難過時，我能給他（她）適當的情緒支持。
	8	我和他（她）心靈能相互溝通契合。
	15	我們彼此能分憂解勞。
	16	我和他（她）已經有親吻的親密舉動。
	18	我覺得要告訴他（她）我的感受是很容易的事。
	22	我和他（她）可說是無所不談。
Commitment	7	我願與他（她）共度一生。
	10	我仍然保持觀望，希望那一天能碰到更好的對象。
	12	我還沒準備好接受和他（她）的這份情感。
	13	我們有各自的路要走，還不到相互承諾的時候。
	14	有時我覺得自己對他（她）還是有點虛情假意，不是很真實。
	17	我還不想那麼快就和他（她）定下來。
	20	對於終身的伴侶，現在的我還不知如何抉擇。
	21	搞不清楚自己是真愛他（她）呢？還是只是習慣兩人在一起？

Note. Items 10, 12, 13, 14, 17, 20, and 21 are negatively worded items.