ESTIMATION OF DIAGNOSTIC CLASSIFICATION MODELS WITHOUT CONSTRAINTS: ISSUES WITH CLASS LABEL SWITCHING

BY

HONGLING LAO

Submitted to the graduate degree program in the Department of Educational Psychology

and

the Graduate Faculty of the University of Kansas in partial fulfillment of the

requirements

for the degree of Master of Science in Education.

Neal Kingston

Advisor & Co-Chair: Neal Kingston

Co-Chair: Jonathan Templin

Committee Member: William Skorupski

Date defended: 6/7/2016

ESTIMATION OF DIAGNOSTIC CLASSIFICATION MODELS WITHOUT CONSTRAINTS: ISSUES WITH CLASS LABEL SWITCHING

BY

HONGLING LAO

Submitted to the graduate degree program in the Department of Educational Psychology

and

the Graduate Faculty of the University of Kansas in partial fulfillment of the

requirements

for the degree of Master of Science in Education.

Neal Kingston

Advisor & Co-Chair: Neal Kingston

Co-Chair: Jonathan Templin

Committee Member: William Skorupski

Date defended: 6/7/2016

Abstract

Diagnostic classification models (DCMs) may suffer from the latent class label switching issue. Label switching refers to the situation where the labels for the parameters switch across replications of the same estimation. It happens when there are the permutations of the number of latent classes (k!) with statistically equivalent solutions to the estimation, resulting from a symmetry parameter space. With uncertainty in the accuracy of the labels in the parameters, the interpretation of results could be invalid and misleading.

A simulation study is used to investigate the prevalence of label switching in DCMs. Three independent variables are involved, including the model constraints, the effect size of the measurement model parameters, and the q-matrix specifications. The data is generated via R, and estimated via Mplus. Label switching is operationally defined as, for the same dataset, the existence of any difference in the estimated parameters between the model without constraints and the model with constraints, given that they have the same log likelihood.

Results show that local optimal solutions prevail in some conditions, making it difficult to identify label switching. Given the same log likelihoods between models, 13.9 to 40 percent of those replications show label switching.

Acknowledgements

None of my current achievements could have happened without my advisor, Dr. Kingston. He recognized my potential from the beginning and introduced me to this intriguing field. Without his strong and generous support for the past few years, I could hardly have learned so much, and built a vision for the future. Neal, you have my greatest gratitude!

I owe my debt to Dr. Templin, who co-chairs my thesis with my advisor. I gained a much deeper understanding of statistics from Dr. Templin's coaching as well as his classes in three years. His detailed feedback on doing research equipped me with independence in future research. I am very thankful for learning these important skills and knowledge from you!

I owe my debt to Dr. Skorupski, who serves as a valuable committee member, providing feedback on the simulation study. I have learned a lot of measurement theories from Dr. Skorupski, as well as the application of Bayesian statistics. I appreciate both your well-organized instructions and the rock-style musical taste at Face-The-Music!

The encouragement from and accompany of my best friends have played a key role in this long process, which demands high self-motivation and constant self-discipline. The moments shared with you have filled my time at this lovely town with a lot of fun, learning and unforgettable memories!

The love and support from my family has been the spiritual guidance for my persistence and resilience. You have nourished my heart and built my soul. I love you. Thank you!

Table of Contents

ABSTRACT-- 3

ACKNOWLEDGEMENTS-- 4

CHAPTER 1: INTRODUCTION --- 7

CHAPTER 2: LITERATURE REVIEW ---11

OVERVIEW OF DCMS---11

DCMS AS LATENT CLASS MODELS---12

THE MEASUREMENT MODELS --13

The Log Linear Cognitive Diagnosis Model---13

Model Constraints in LCDM and Label Switching ---18

THE STRUCTURAL MODELS---19

The Log Linear Saturated Structural Models---20

AN EXAMPLE OF LABEL SWITCHING IN LATENT CLASS MODELS-----------------------------------24

CHAPTER 3: METHOD---28

CHAPTER 4: RESULTS---31

BALANCED Q-MATRIX --31

EMPIRICAL Q-MATRIX --33

PARAMETER RECOVERY---34

CHAPTER 5: DISCUSSION ---38

REFERENCES ---42

List of Tables and Figures

Table 1 *An Example of a Q-matrix for Three Attributes* 45

Table 2 *An Example of Attribute Profiles and Probabilities for Three Attributes* 46

Table 3 *An Example of Label Switching with Macready and Dayton (1977) Data* 47

Table 4 *The Balanced and the Empirical Q-Matrix Specification* 48

Table 5 *The Structural Model Parameter Specification in the Simulation Study*......................... 50

Table 6 The *Measurement Model Parameter Specification in the Simulation Study*................... 51

Table 7 *Main Results for the Balanced Q-Matrix and Empirical Q-Matrix*................................ 52

Table 8 *Parameter Recovery Results for the Structural Model Parameters* 53

Table 9 *Parameter Recovery Results for the Measurement Model Parameters*........................... 54

Table 10 *Parameter Recovery Results for the Person Latent Class Classifications* 56

Table 11 *Parameter Recovery Results for the Structural Model Parameters between the Non-Label Switching and the Label Switching Replications*... 57

Table 12 *Parameter Recovery Results for the Measurement Model Parameters for the Non-Label Switching Replications*... 59

Table 13 *Parameter Recovery Results for the Measurement Model Parameters for the Label Switching Replications*... 61

Table 14 *Parameter Recovery Results for the Person Latent Class Classifications between the Non-Label Switching and the Label Switching Replications*.. 63

Figure 1 *Main Results for the Balanced Q-Matrix Condition* 64

Figure 2 *Main Results for the Empirical Q-Matrix Condition* 65

Chapter 1: Introduction

This study targets an estimation issue in diagnostic classification models (DCMs) that occurs when they are estimated without monotonicity constraints. DCMs, as the name implies, are specific kinds of measurement models that classify people so as to make diagnostic decisions, e.g., student placements in education, or psychological pathological diagnoses. Statistically, DCMs are confirmatory latent class models. Latent class models, introduced by Lazarsfeld and Henry (1968) and further developed by Goodman (1974a, 1974b), are intended to build a relationship between unobserved (latent) categorical variables (classes) and observed (manifest) variables (Kaplan, 2004).

Label switching in latent class models (Redner and Walker, 1984), happens when the label for each latent class switches across replicated estimations. For example, for a two-class model, at the first estimation, class 1 is labeled as having an emotional disorder. Accordingly, class 2 is labeled as having no emotional disorder. However, for another estimation with the same model specification and using the same software, unexpectedly, class 1 is now labeled as having no emotional disorder, whereas class 2 is labeled as having an emotional disorder. This is label switching. With uncertainty in the accuracy of the labels in the parameters, the interpretation of results could be invalid and misleading.

For *exploratory* latent class models the number of latent classes is unknown, so is the meaning of each latent class. However, for the *confirmatory* latent class models, both the number and the meaning of each latent class is predefined, such as those in DCMs. As a result, label switching is a concern in *confirmatory* latent class models (the focus of this paper), but irrelevant in *exploratory* models.

Statistically, label switching is due to the fact that there is more than one set of parameters that produces the same model likelihood. The number of statistically equivalent alternative solutions is equal to the number of permutations of the count of classes. With K latent classes, there are K! statistically equivalent solutions resulting from symmetry of the likelihood function in the parameter space. In such solutions, the values of the parameters are identical within a class, but the class itself may appear with a different numeric label from what was originally assigned. All the permutations of the parameters yield invariant estimation results—each has the same log-likelihood value.

The term “label switching” comes from the fact that the labels of classes (and subsequently, each class-specific parameter), along with their corresponding meaning, can switch across replicated model estimations. Moreover, for exploratory latent class analyses, the label switching issue is not obvious unless comparing results across multiple analyses. A review of the existing methods on the label switching issue is beyond the scope of this paper; please refer to other authors for details (e.g., Celeux, Hurn, & Robert, 2000; Grün & Leisch, 2009; Jasra, Holmes, & Stephens, 2005; Richardson & Green, 1997; Stephens, 2000; Yao, 2012, 2013).

This is a concern in *confirmatory* latent class models (e.g., DCMs), but not to the *exploratory* latent class models because for the *exploratory* models, neither the number nor the meaning of each latent class is defined in the model. They are meant to be provided after the estimation. Consequently, whether the label of having emotional disorder switches from class 1 to class 2 makes no difference, as the meaning (the label) of classes is unknown before estimation. Exploratory models do not assume a fixed label across replicated estimation.

However, *confirmatory* latent class models do assume a consistent labeling across replicated estimations. The label, as well as its corresponding meaning, is predefined for each specific class during the model specification. For example, class 1 is defined explicitly as having emotional disorder in the model. However, inspection of item responses might show that class 1 is associated with items that indicate a lack of an emotional disorder, contradicting the original predefined label. Based on the label in the model specification, those classified in class 1, which are thought to have emotional disorder, are in fact the opposite in the results. The output label is inconsistent with their original definition in the model, which may cause misleading interpretation of the results.

Then the next question is does the process of setting a confirmatory model, either explicitly or implicitly, automatically constrain the labels as well? The answer is not known definitively as it depends on the type of model specification. For example, with four different structural models in DCMs, which will be introduced later in the paper, some may fix the labels as part of the model specification, while others may not. This is the part of the question this paper intends to address.

When not provided with model-specific constraints, DCMs, as confirmatory latent class models, may suffer from this latent class model label switching issue. In DCMs, each latent class represents a specific pattern of mastered attributes. As DCM item and structural parameters are formed from their constrained latent class model counterparts, the meaning of the class is needed to determine which of these latent class model parameters becomes part of the DCMs' parameterization. Because of this *a priori* assignment in DCM analysis an instance of label switching may lead to inconsistent results.

General versions of DCMs, such as the Loglinear Cognitive Diagnosis Model (LCDM; Henson, Templin, & Willse, 2009), by definition place constraints on the item parameters of the model to ensure monotonicity of item response performance with respect to attribute status.

When one places no constraints on model parameters, the label switching issue may jeopardize the meanings of parameters and the interpretations of estimation results.

Despite this risk, not placing constraints on LCDM item parameters is common in nearly all currently available software (e.g., flexMIRT, CDM package in R), possibly due to a lack of widespread recognition of the label switching issue in the community of researchers and practitioners of DCMs. In order to improve the accuracy of using DCMs this thesis will use a simulation study to explore the prevalence of label switching issue in DCMs.

The results will shed light on (1) whether the practitioners of DCMs need to explicitly place model constraints on the parameters in order to reduce the risk and confusion caused by label switching, and (2) whether the current and future software developers of DCMs should add model constraints to the model specification as a common practice in the field.

Chapter 2: Literature Review

Overview of DCMs

DCMs are confirmatory latent class models that provide information for diagnostic decision making (Rupp, Templin, and Henson, 2010). Similar to Item Response Theory (IRT) models, the analytic unit in DCMs is an item. Different from the continuous latent trait assumed in IRT models, DCMs assume categorical latent variables that relate to observed item responses. The unobserved categorical latent variables in DCMs are called attributes. Most often an attribute has two categories and can either be mastered or non-mastered. DCMs provide a profile of mastered attributes to respondents instead of a single score.

DCMs provide two types of information: structural information and item information. Structural information refers to the probability of diagnoses, often called the structural model in DCMs. Item information indicates how well each item discriminates between respondents with different levels of the attribute, and is often called the measurement model in DCMs.

One benefit of using DCMs is that, if the model fits, attribute estimates can have overall higher reliability with fewer items than IRT models, as they require less information to accurately categorize test-takers than it does to place them on a continuous scale (Templin & Bradshaw, 2013). Therefore, tests developed based on DCMs can assess multiple attributes simultaneously without as many high dimensional calculations as are required in multidimensional IRT models. The price for this multidimensional efficiency is the reduction of information obtained for each attribute. The classification in DCMs comes from the latent variables themselves as opposed to having to find a cut score on the estimated latent variable dimension. As such, DCMs are appropriate for diagnostic assessments, including clinical (e.g., Templin & Henson, 2006) and psychoeducational assessments (e.g., Rupp & Templin, 2008).

DCMs as Latent Class Models

To better understand the label switching issue, it may be beneficial to provide a background on the statistical foundations of DCM. Since DCMs are constrained latent class models, it is beneficial to begin with latent class models. A latent class model includes both structural and the measurement components. The structural models describe the relationship among attributes. The measurement models specify the relationship between item responses and attribute mastery status. The full model can be mathematically expressed in Equation (1), representing the marginal likelihood function for binary items with binary attribute for a given respondent.

$$\mathrm{P}(\mathbf{X}_{\mathrm{r}}) = \sum_{\mathrm{c}=1}^{2^{A}} \nu_{\mathrm{c}} \prod_{\mathrm{i}=1}^{\mathrm{I}} \mathrm{P}(\mathrm{X}_{\mathrm{ri}} = 1|\boldsymbol{\alpha}_{\mathbf{rc}})^{\mathrm{x}_{\mathrm{ri}}} \left(1 - \mathrm{P}(\mathrm{X}_{\mathrm{ri}} = 1|\boldsymbol{\alpha}_{\mathbf{rc}})\right)^{1-\mathrm{x}_{\mathrm{ir}}} \tag{1}$$

where $\mathrm{P}(\boldsymbol{X_r})$ is the probability function; $\boldsymbol{X_r}$ are the observed item responses for respondent r; Σ is the summation symbol to sum across 2^{A} latent classes; ν_c is the probability (i.e., marginal frequency) of latent class c in the population; ∏ is the product symbol, multiplying the conditional probabilities of correct response across all items given each latent class c for respondent r; $P(X_{ri} = 1|\boldsymbol{\alpha_{rc}})$ is the conditional probability of correct response to item i by respondent r, given a specific attribute profile c; x_{ri} is the binary item response to item i for respondent r (0/1 for incorrect/correct).

For each respondent, the latent class model searches the latent class space one-by-one, modeling the item responses given each latent class. The probabilities for all respondents are multiplied as a whole for parameter estimation with all the observed data.

DCMs are confirmatory latent class models, meaning DCMs include the two components of latent class models: measurement and structural. For the measurement component, it can be modeled using the general LCDM framework (or with any other latent-class based DCM). For the structural component, there are more options, including the saturated log linear structural model, the reduced log linear structural model, the tetrachoric structural model, and the Bayesian structural model. The measurement models will be introduced first, familiarizing readers with the log linear modeling framework. The structural models follow, with a higher level of complexity.

The Measurement Models

The measurement model in DCMs indicates how well the items discriminate between respondents with different levels of mastery of the attributes. In DCMs, the q-matrix is used to represents the set of attributes each item measures. The q-matrix stores the information in a matrix, with items in the rows and the attributes in the columns. The entries in the matrix are either 0s or 1s, representing whether an item measures an attribute (i.e., whether subject matter experts believe that the cognitive process of responding to an item involves a particular attribute). Statistically, it is analogous to the loading matrix in confirmatory factor analysis models, specifying an empirical hypothesis about the attribute structure to be tested.

The Log Linear Cognitive Diagnosis Model

Statistical Expression

In DCMs, the response to an item is assumed to be caused by a function of only the set of attributes measured by the item. Given an examinee’s attribute profile and item parameters, item responses are conditionally independent. An attribute profile in DCMs, which will be described in more detail in the structural models, represents a latent class (i.e., a specific combination of mastery statuses for all attributes).

As an item is developed to measure a set of attributes, the mastery status of the corresponding attributes should be a good predictor of the response to the item. If an item is not designed to measure an attribute, the mastery status of that attribute is assumed to have no effect on that item response. With an increasing number of mastered measured attributes, the predicted probability of a correct response to the item increases, as well as the log-odds of a correct response.

For example, take an item measuring two attributes. The probability of a correct response to an item can be modeled with the LCDM. The model uses concepts borrowed from the analysis of variance (ANOVA) model, and breaks down item response probabilities into model-based parameters, such as an intercept, main effects, and interactions. The intercept indicates the probability of a correct response to the item when no attributes are mastered. A main effect represents the difference in probability between those who have mastered the attribute and those who have not, conditional on non-mastering the other attribute. An interaction term is used to explain any increase in the probability for those who have mastered both attributes, over and above the sum of two main effects.

To solve the issue of constrained range, a logit link function is used, so that the new model space ranges from negative infinity to positive infinity. The logit is the natural logarithm of the odds. The odds is the ratio of the probability of a correct response to an item over the probability of an incorrect one.

The general equation for items measuring two attributes is as following Equation (2):

$$\ln\left(\frac{\mathrm{P}(\mathrm{X}_{\mathrm{ie}}=1|\alpha_{\mathrm{e}1},\alpha_{\mathrm{e}2})}{\mathrm{P}(\mathrm{X}_{\mathrm{ie}}=0|\alpha_{\mathrm{e}1},\alpha_{\mathrm{e}2})}\right)=\lambda_{\mathrm{i},0}+\lambda_{\mathrm{i},1,(1)}\alpha_{\mathrm{e}1}+\lambda_{\mathrm{i},1,(2)}\alpha_{\mathrm{e}2}+\lambda_{\mathrm{i},2,(1,2)}\alpha_{\mathrm{e}1}\alpha_{\mathrm{e}2} \qquad (2)$$

where, X_{ie} is the response to the item i for examinee e; $X_{ie} \in \{0,1\}$ indicates a correct/incorrect response to item i; α_{e1} is the latent variable for Attribute 1; $\alpha_{e2}$is the latent variable for Attribute 2; $\mathrm{P}(\mathrm{X}_{\mathrm{ie}}=1|\alpha_{\mathrm{e}1},\alpha_{\mathrm{e}2})$ rrepresents the probability of a correct response to item i for an examinee e, given the mastery status (0/1) of attribute α_1 and α_2; $\mathrm{P}(X_{ie}=0 \mid \alpha_{\mathrm{e}1},\alpha_{\mathrm{e}2})$ represents the probability of an incorrect response to an item i, given the mastery status of α_{e1} and α_{e2}; $\ln\left(\frac{\mathrm{P}(\mathrm{X}_{ie}=1|\alpha_{\mathrm{e}1},\alpha_{\mathrm{e}2})}{\mathrm{P}(\mathrm{X}_{ie}=0|\alpha_{\mathrm{e}1},\alpha_{\mathrm{e}2})}\right)$ represents the log-odds of a correct response over an incorrect response to an item i, given the mastery status of α_{e1} and α_{e2}; $\lambda_{i,0}$ is the intercept and represents the predicted log-odds of a correct response for examinees in the reference group, who have mastered neither of the attributes (i.e., α_{e1}= 0, α_{e2} = 0); $\lambda_{i,1,(1)}$ is the main effect for α_{e1}, indicating the increase in the log-odds of a correct response for examinees who have mastered α_{e1} but not α_{e2} (i.e., α_{e1}= 1, α_{e2}= 0); $\lambda_{i,1,(2)}$ is the main effect for α_{e2}, indicating the increase in the log-odds of a correct response for examinees who have mastered α_{e2} but not α_{e1} (α_{e1}= 0, α_{e2}= 1); $\lambda_{i,2,(1,2)}$ is the interaction for α_{e1} and α_{e2}, indicating the additional increase in the log-odds for examinees who have mastered both α_{e1} and α_{e2} (i.e., α_{e1} = 1, α_{e2} = 1).

The log-odds can be transformed back into the probability of correct response with Equation (3).

$$\mathrm{P}(\mathrm{X}_{\mathrm{ie}}=1|\alpha_{\mathrm{e}1},\alpha_{\mathrm{e}2})=\frac{\exp\left(\lambda_{\mathrm{i},0}+\lambda_{\mathrm{i},1,(1)}\alpha_{e1}+\lambda_{\mathrm{i},1,(2)}\alpha_{\mathrm{e}2}+\lambda_{\mathrm{i},2,(1,2)}\alpha_{\mathrm{e}1}\alpha_{\mathrm{e}2}\right)}{1+\exp\left(\lambda_{\mathrm{i},0}+\lambda_{\mathrm{i},1,(1)}\alpha_{e1}+\lambda_{\mathrm{i},1,(2)}\alpha_{\mathrm{e}2}+\lambda_{\mathrm{i},2,(1,2)}\alpha_{\mathrm{e}1}\alpha_{\mathrm{e}2}\right)} \qquad (3)$$

For items measuring multiple attributes simultaneously, a generalized form of the log linear model can be used, as shown in Equation ((4). It specifies an item response as a function of the q-matrix entries ($\boldsymbol{q_i}$, the attribute profiles ($\boldsymbol{\alpha_c}$), and the item parameters ($\boldsymbol{\lambda_i}$).

$$\pi_{ic} = P(X_{ic} = 1|\boldsymbol{\alpha_c}) = \frac{\exp(\lambda_{i,0} + \boldsymbol{\lambda_i^T h(\alpha_c, q_i)})}{1 + \exp(\lambda_{i,0} + \boldsymbol{\lambda_i^T h(\alpha_c, q_i)})} \tag{4}$$

Where π_{ic} is the probability of a correct response to an item i ($P(X_{ri} = 1|\boldsymbol{\alpha_{rc}})$) from Equation (1), given latent class c; $\lambda_{i,0}$ is the intercept parameter, indicating the logit of a correct response given none of the attributes measured by item i is mastered; $\boldsymbol{\lambda_i}$ are the main effect and interaction parameters for item i, a vector of size $(2^n - 1) \times 1$; $\boldsymbol{h(\alpha_c, q_i)}$ is a vector of the same size with $\boldsymbol{\lambda_i}$, pertaining to the linear combinations of $\boldsymbol{\alpha_c}$and $\boldsymbol{q_i}$, where $\boldsymbol{\alpha_c}$ represents the attribute profile in latent class c, and $\boldsymbol{q_i}$, represents the q-matrix entries for item i.

The exponent part is a kernel expression, which can be represented in Equation (5). The kernel expression includes an intercept, all main effects, and all possible interactions between attributes.

$$\begin{aligned} \text{kernel} &= \lambda_{i,0} + \boldsymbol{\lambda_i^T h(\alpha_c, q_i)} \\ &= \lambda_{i,0} + \sum_{a=1}^{A} \lambda_{i,1,(a)} \alpha_{ca} q_{ia} + \sum_{a=1} \sum\nolimits_{a'>1}^{A} \lambda_{i,2,(a,a')} \alpha_{ca} \alpha_{ca'} q_{ca} q_{ca'} \\ &+ \ldots \end{aligned} \tag{5}$$

where $\lambda_{i,0}$ is the intercept parameter; $\lambda_{i,1,(a)}$ represent the main effect for attribute α; α_{ca} is the attribute profile for latent class c; q_{ic} is the q-matrix entry for item i in latent class c; $\lambda_{i,2,(a,a')}$ represents three two-way interactions between two attributes; etc.

An Example with Three Attributes

To facilitate understanding, an example with three attributes and seven items is shown. With three attributes, there are seven possible unique patterns on the q-matrix, as shown in Table 1. The number of unique patterns for the q-matrix is $(2^A - 1)$, where A is the number of binary attributes in the test. There are 2^A unique patterns. Yet an item must measure at least one attribute, so he pattern in which no attributes is measured is excluded.

Presumably, the mastery status of an attribute should not affect the probability of a correct response to an item if the item does not measure the attribute; vice versa. For example, for item 1, the probability of a correct response will only be affected by the mastery status of attribute 3, whereas the mastery status of attribute 1 or attribute 2 has no effect. For item 7, the mastery status of all three attributes will have an effect on the probability of a correct response.

Based on Equation ((4) and (5), given an attribute profile (i.e., latent class c), the probability of correct response to item i is expressed as following Equation (6).

$$\begin{aligned} \pi_{ic} = \mathrm{P}(\mathrm{X}_{ic} = 1|\boldsymbol{\alpha}_\mathbf{c}) &= \frac{\exp(\lambda_{i,0} + \boldsymbol{\lambda}_\mathbf{i}^\mathbf{T}\mathbf{h}(\boldsymbol{\alpha}_\mathbf{c}, \mathbf{q}_\mathbf{i}))}{1 + \exp(\lambda_{i,0} + \boldsymbol{\lambda}_\mathbf{i}^\mathbf{T}\mathbf{h}(\boldsymbol{\alpha}_\mathbf{c}, \mathbf{q}_\mathbf{i}))}; \\ \textit{where } \boldsymbol{\lambda}_\mathbf{i}^\mathbf{T}\mathbf{h}(\boldsymbol{\alpha}_\mathbf{c}, \mathbf{q}_\mathbf{i}) &= (\lambda_{i,1,1})(\alpha_{c1})(q_{i1}) + (\lambda_{i,1,2})(\alpha_{c2})(q_{i2}) + (\lambda_{i,1,3})(\alpha_{c3})(q_{i3}) \\ &+ (\lambda_{i,2,(1,2)})(\alpha_{c1})(\alpha_{c2})(q_{i1})(q_{i2}) \\ &+ (\lambda_{i,2,(1,3)})(\alpha_{c1})(\alpha_{c3})(q_{i1})(q_{i3}) \\ &+ (\lambda_{i,2,(2,3)})(\alpha_{c2})(\alpha_{c3})(q_{i2})(q_{i3}) \\ &+ (\lambda_{i,3,(1,2,3)})(\alpha_{c1})(\alpha_{c2})(\alpha_{c3})(q_{i1})(q_{i2})(q_{i3}) \end{aligned} \tag{6}$$

where exp(·) is the exponential function; λ is the item parameter, including intercept $(\lambda_{i,0})$, main effects$(\lambda_{i,i,1})$, two-way interactions $(\lambda_{i,2,(a,a')})$, and three-way interactions $\left(\lambda_{i,3,(1,2,3)}\right)$; α_{ca} is the attribute profile, indicating the mastery status of attribute a for latent class c; q_{ic} is the binary q-matrix entry, indicating whether item i measures the attribute a.

As an example, the probability of a correct response to item 5, given a person in latent class c is $\pi_{5c,}$ representing in Equation (7).

$$\begin{aligned}
\pi_{5c} = \; & P(X_{5c} = 1|\boldsymbol{\alpha}_{\mathbf{c}}) = \frac{\exp(\lambda_{5,0} + \boldsymbol{\lambda}_{\mathbf{5}}^{\mathbf{T}}\mathbf{h}(\boldsymbol{\alpha}_{\mathbf{c}}, \mathbf{q}_{\mathbf{5}}))}{1 + \exp(\lambda_{5,0} + \boldsymbol{\lambda}_{\mathbf{5}}^{\mathbf{T}}\mathbf{h}(\boldsymbol{\alpha}_{\mathbf{c}}, \mathbf{q}_{\mathbf{5}}))}; \\
& where\ \boldsymbol{\lambda}_{\mathbf{5}}^{\mathbf{T}}\mathbf{h}(\boldsymbol{\alpha}_{\mathbf{c}}, \mathbf{q}_{\mathbf{5}}) = (\lambda_{5,1,1})(\alpha_{c1})(q_{51}) + (\lambda_{5,1,2})(\alpha_{c2})(q_{52}) \\
& + (\lambda_{5,1,3})(\alpha_{c3})(q_{53}) + (\lambda_{5,2,(1,2)})(\alpha_{c1})(\alpha_{c2})(q_{51})(q_{52}) \\
& + (\lambda_{5,2,(1,3)})(\alpha_{c1})(\alpha_{c3})(q_{51})(q_{53}) \\
& + (\lambda_{5,2,(2,3)})(\alpha_{c2})(\alpha_{c3})(q_{52})(q_{53}) \\
& + (\lambda_{5,2,(1,2,3)})(\alpha_{c1})(\alpha_{c2})(\alpha_{c3})(q_{51})(q_{52})(q_{53}) \qquad (7) \\
& = (\lambda_{5,1,1})(\alpha_{c1})(1) + (\lambda_{5,1,2})(\alpha_{c2})(0) + (\lambda_{5,1,3})(\alpha_{c3})(1) \\
& + (\lambda_{5,2,(1,2)})(\alpha_{c1})(\alpha_{c2})(1)(0) + (\lambda_{5,2,(1,3)})(\alpha_{c1})(\alpha_{c3})(1)(1) \\
& + (\lambda_{5,2,(2,3)})(\alpha_{c2})(\alpha_{c3})(0)(1) \\
& + (\lambda_{5,2,(1,2,3)})(\alpha_{c1})(\alpha_{c2})(\alpha_{c3})(1)(0)(1) \\
& = (\lambda_{5,1,1})(\alpha_{c1}) + (\lambda_{5,1,3})(\alpha_{c3}) + (\lambda_{5,2,(1,3)})(\alpha_{c1})(\alpha_{c3})
\end{aligned}$$

Therefore, the response to item 5 is only affected by the mastery status of attributes 1 and 3.

Model Constraints in LCDM and Label Switching

A brief introduction to the constraints used in the LCDM will facilitate understanding of the effects such constraints have on label switching.

In the LCDM, the main effect parameters, representing the logit probability difference of a correct response to an item between masters and non-masters of measured attributes, are usually constrained to be greater than zero to meet the assumption of a monotonic increase in the probability with increasing mastery: with an increasing number of mastered measured attributes, students should have higher probability of a correct response. Secondly, constraints can be placed on the interaction effect parameters so as to ensure the simultaneous mastery of more than one measured attribute will result in a higher probability of a correct response than mastery of any attribute alone.

With the constraints on main effect parameters, the meaning of latent classes is fixed to match the model specification. Without model such constraints, the meaning of latent classes is determined during the estimation. With model constraints, the previous unspecified parameter space that leads to the label switching is resolved. However, in addition to constraints on main effect parameters, it is not yet understood whether constraints on interaction effect parameters are necessary to prevent label switching.

The Structural Models

In DCMs, the structural model provides the probability of diagnoses (i.e., proportion population mastering of each attribute profile). Assuming there are Abinary attributes (i.e., mastery/non-mastery), there are 2^A permutations of attribute mastery statuses, called attribute profiles. As DCMs are constrained latent class models, each attribute profile represents a latent class. The probability of all possible attribute profiles sums up to one. For example, with three measured attributes($A = 3$), there are $2^3 = 8$ different attribute profiles of mastery or non-mastery for these three attributes. The attribute profile that represents none of the three attributes mastered is symbolized as 000, whereas it is 001 if the first three attributes non-mastered and the last one mastered. The sum of eight probabilities equal one. As such, there are (2^A – 1) unknown probabilities in the structural model. Thus seven probabilities need to be estimated and the eighth can be calculated by subtracting the sum of the others from one. This model is often called the saturated structural model as all possible structural probabilities are estimated, making it equivalent to the general latent class structural model.

One potential issue in estimation is that the number of unknown parameters in the saturated structural model increases exponentially as the number of measured attributes increases. Thus except for the saturated structural models, there are various constrained models that aim to reduce the number of unknown parameters to improve estimation efficiency. To achieve this reduction, the marginal and joint probabilities of attributes can be modeled as various kinds of effects. For example, in the log linear structural model, the joint probabilities can be represented by main effects and interaction terms, instead of estimating all $2^A - 1$ marginal probabilities linearly. The following section will only introduce the general saturated model, as the constrained model is beyond the scope of the current study.

The Log Linear Saturated Structural Models

Statistical Expression

Log linear models re-express the joint probability into different effects similar to ANOVA models, which provides a model-based method to reduce redundant parameters (see Chapter 8 of Rupp et al., 2010). A log linear model rescales the base-rate probability (ν_c) for each attribute profile (latent class) c using the natural logarithm ($\mu_c = \log \nu_c$) . The attribute profile probabilities can be integrated into the marginal probabilities of mastery for each attribute—the proportion of examinees mastering any specific attribute.

The log linear saturated model uses main effects for attributes and possible interactions of all attributes to express the log of the base-rate probability for a given attribute profile. It can be represented as below in Equation (8).

$$\mu_c = \sum_{a=1}^{A} \Upsilon_{1,(a)}\alpha_{ca} + \sum_{a=1}^{A-1}\sum_{a'=a+1}^{A} \gamma_{2,(a,a')}\alpha_{ca}\alpha_{ca'} + \cdots + \gamma_{A,(a,a',\ldots)}\prod_{a=1}^{A}\alpha_a \qquad (8)$$

where μ_c is the kernel expression to indicate the membership probability of latent class c; $\Upsilon_{1,(a)}$represents the main effect connected with attribute a, indicating the mastery probability in a natural logarithmic scale of each single attribute a, whereas in the measurement model, the main effect parameters indicate the difference of a correct-response probability to an item in logit scale for those who have mastered a single attribute, compared with those who have mastered none) ; $\gamma_{2,(a,a')}$represents the two-way interaction connected with attribute a and a', indicating the mastery probability in a natural logarithmic scale of both attributes, whereas in the measurement mode, two-way interaction parameters indicate the difference of a correct-response probability to an item in logit scale for those who have mastered both attributes, compared with those who have mastered either one; $\gamma_{A,(a,a',\ldots)}$represents a one-way interaction connected with all attributes, indicating the mastery probability in a natural logarithmic scale of all a attributes; α_{ca} represents the attribute profile in latent class c.

The interpretation of parameters in the log linear structural model is not the same as that in the log linear measurement model. This is because in the structural model, parameters are used to estimate the relationship between attributes, whereas parameters are used to estimate the relationship between attributes and the items in the measurement model. These parameters are designed to answer different questions, and therefore carry different meanings. The structural model depends on the probability (proportion) of mastering each attribute in the population (i.e., estimation of the distribution of attribute mastery status), whereas the measurement model depends on the probability of a correct response to an item (i.e., estimation of the relationship between attribute mastery status and item property).

Based on the parameters in equation (8), the probability of membership in latent class c (ν_c) can be estimated with Equation (9). The probability can be calculated by an inverse transformation of a natural logarithm scale ($\exp(\mu_c)$) . Next, in order to ensure that the sum of all probabilities is equal to one, each transformed component ($\exp(\mu_c)$) is divided by the sum of all components, as shown in Equation (9).

$$\mathrm{v}_c = \frac{\exp(\mu_c)}{\sum_{c=1}^{c}\exp(\mathrm{\mu}_c)} \tag{9}$$

Furthermore, the marginal probability of each latent class can be aggregated into the marginal probability (p_a) for each attribute as shown in Equation (10),

$$\mathrm{p}_a = \sum_{c=1}^{c} \alpha_{ca}\mathrm{v}_c \tag{10}$$

where v_c is the probability of membership in latent class c; α_{ca} is binary (0/1) variable, representing whether Attribute a is non-mastered (0) or mastered (1) in latent class c.

An Example with Three Attributes

To facilitate understanding, an example with three attributes is shown. With three attributes, there are $2^3 = 8$ unique attribute profiles (i.e., latent classes). The corresponding attribute profile and probabilities are shown in Table 2. There are two types of marginal probabilities: one is for each attribute (p_a), the other is for each latent class (ν_c). For the marginal probabilities of latent classes, they are re-expressed in a linear combination of attributes in the model-based combinations. Given that the probability can only range between zero and one, it is rescaled in the natural logarithm (μ_c) so that the model space can be wider and more flexible.

In the log linear saturated model, the log probability for latent class c (μ_c) is equal the sum of three main effects, three two-way interactions, and one three-way interaction. This is demonstrated using latent Class 5 as an example in Equation (11),

$$\begin{aligned}\mu_5 = {} & (\gamma_{1,1})(\alpha_{51}) + (\gamma_{1,2})(\alpha_{52}) + (\gamma_{1,3})(\alpha_{53}) + (\gamma_{2,(1,2)})(\alpha_{51})(\alpha_{52}) \\ & + (\gamma_{2,(1,3)})(\alpha_{51})(\alpha_{53}) + (\gamma_{2,(2,3)})(\alpha_{52})(\alpha_{53}) \\ & + (\gamma_{3,(1,2,3)})(\alpha_{51})(\alpha_{52})(\alpha_{53}) \\ & = (\gamma_{1,1})(1) + (\gamma_{1,2})(0) + (\gamma_{1,3})(0) + (\gamma_{2,(1,2)})(1)(0) \\ & + (\gamma_{2,(1,3)})(1)(0) + (\gamma_{2,(2,3)})(0)(0) + (\gamma_{3,(1,2,3)})(1)(0)(0) \\ & = (\gamma_{1,1})\end{aligned} \tag{11}$$

in which μ_5 is the probability for latent class 5; $\gamma_{1,1}$is the main effect parameter for attribute 1; $\gamma_{2,(1,2)}$ is the two-way interaction between attribute 1 and attribute 2; $\gamma_{3,(1,2,3)}$ is the three-way interaction among attribute 1, attribute 2, and attribute 3; and $\alpha_{51}$is the latent variable for attribute 1 in latent class 5.

The log probability for latent class can be transformed back to a probability scale using Equation (9). The transformation involves two steps. The first step is to calculate the exponential of log probability, which is the raw probability that may not fall within the range of zero and one.

In order to ensure that the sum of probabilities is equal to one, each raw probability is divided by the sum of all raw probabilities.

Using latent class 5 as an example, as shown in Equation (12),

$$\mathrm{\nu}_5 = \frac{\exp(\mathrm{\mu}_5)}{\sum_{c=1}^{c=5}\exp(\mathrm{\mu}_c)} = \frac{\mathrm{\nu}_5{}'}{\mathrm{\nu}_1{}' + \mathrm{\nu}_2{}' + \mathrm{\nu}_3{}' + \mathrm{\nu}_4{}' + \mathrm{\nu}_5{}'} \tag{12}$$

in which ν_c' the log probability of latent class c; $\mathrm{\nu}_5$ is the marginal probability of latent class 5.

Finally, these probabilities for latent classes ν_c can be used to calculated marginal probabilities for attributes. The probability of attribute 1 is shown as an example in Equation (13).

$$\mathrm{p}_1 = \sum_{c=1}^{c} \alpha_{c1}\mathrm{\nu}_c$$

$$= (0)(\nu_1) + (0)(\nu_2) + (0)(\nu_3) + (0)(\nu_4) + (1)(\nu_5) + (1)(\nu_6) \tag{13}$$

$$+ (1)(\nu_7) + (1)(\nu_8) = (\nu_5) + (\nu_6) + (\nu_7) + (\nu_8)$$

In summary, the log probabilities of latent classes $\mathrm{\mu}_c$are modeled by main effect and interaction effect parameters γ (Equation (8)). The log probabilities $\mathrm{\mu}_c$are then transformed back to probability scale (Equation (9)). Finally, the probabilities of latent class ($\mathrm{\nu}_c$) are used to calculate the marginal probabilities of attributes p_a (Equation (10)).

An Example of Label Switching in Latent Class Models

A concrete example may assist in building an intuitive understanding of the label switching issue. With a single attribute measured, there are two latent classes (non-mastery/mastery of the attribute), and thus there is only one way of label switching between the two classes (0 or 1). In this case, it is easy to judge which version of the results is the "correct" one, based on the monotonicity assumption. That is, students mastering the attribute have higher probability of a correct response to the items than those non-mastered. In other words, the main effect parameters have to be positive, representing the probability increment (in logit scale) of a correct response to an item, between the students with mastery and those without.

Nevertheless, it becomes increasingly challenging when more attributes are measured. With *A* attributes measured, the number of latent classes increase exponentially(2^A). Correspondingly, the number of permutations of possible label switching increases rapidly$(2^A!)$. In this case, it is difficult to identify and disentangle the label switching pattern and recover the "correct" labeling as defined initially in the model, and thus misinterpretation becomes much more likely. Strategically, it is much more practical to avoid the issue from the beginning than to recover from any unforeseen consequences after they occur.

The sample data generated is based on the observed frequency of response patterns in Table 3 $\mathrm{D}_2$from Macready and Dayton (1977). A simple dataset such as that shown in Table 3 can be used to demonstrate this effect in a straightforward manner. In total, there is one attribute, four items, and 142 binary responses. A one-attribute DCM is used to analyze the data, with the specification of the log linear structural model and LCDM measurement model, using RStudio (Version 0.99.473), a tool designed to facilitate the use of R. Even with only two replications, the label switched. The results are shown in Table 3.

With positive main effect parameters, *Replication I* in Table 3 shows the correct version of labeling, while *Replication II* shows the incorrect one after label switching. Most importantly, *Replication II* shares the same model fit index with *Replication I*, with the log likelihood being -

331.76. Label switching happens without influencing the model fit, which makes it a less discernible issue.

For the structural model, the probability of the mastery switches with that of non-mastery after the label switches. The meanings of the parameters switch correspondingly with label switching, yet the labels carry over as the model had originally specified. This is a problem in confirmatory models.

For the measurement model, the main effect parameters in *Replication II* have the same absolute values with those in *Replication I*, with opposite sign. In *Replication I*, the main effect parameters indicates the probability difference on a logit scale of a correct response of class 2 from that of class 1 (the reference group). The positive parameters indicate a higher probability for membership in class 2. As the monotonicity assumption indicates, class 2 should be comprised of students with mastery. On the contrary, in *Replication II*, the main effect parameters with the same meaning, are reversed and have a negative sign. In this case, class 2 actually refers to students without mastery. However, with the same model specification, the label in class 1 is fixed across replication as either mastery or non-mastery. In either case, there is one version of results that is contradictory to what the label indicates. This is the type of misleading information caused by the label switching issue.

While the absolute probability difference between groups stays the same (i.e., same absolute values in main effect parameters with an opposite sign), the probability of the reference group changes, as the group membership switches between two result versions. The actual meaning of the reference group switches between mastery and non-mastery. Accordingly, the intercept parameters, referring to the probability in logit scale of a correct response for the reference group, are different in the two result versions.

We propose to solve this issue by setting the same labeling rules for both processes—model specification and model estimation, so as not to allow random assignment. In the DCMs, this problem caused label switched, but could be resolved by add a constraint such that the probability for a correct response for those students with more mastery monotonically increasing. The occurrence of label switching can be identified by comparing results from the model with redefined labeling rules (i.e., model constraints) with the model without such constraints.

Chapter 3: Method

The following simulation study was conducted to investigate the prevalence of the label switching issue in DCMs and the effectiveness of adding model constraints to resolve the issue. The presence of label switching is the dependent variable, operationally identified as any difference in the structural model parameters between the model with constraints and the model without constraints for the same data set, given that the two models have the same log likelihood.

There are three independent variables. The first one is the use of constraints or the absence of constraints in the DCM measurement models. Based on the monotonic assumption, the constraints restrict the main effect parameters in the measurement models to be greater than zero (pp. 208 - 211, Rupp, Templin & Henson, 2010). The interaction terms are constrained to ensure that the main effect parameters are positive. From a substantive perspective, this means that there is always an increase in the probability of a success response to an item, with more mastery of the attributes measured by an item. Constraints are placed on the measurement models because they connect between latent classes and item performances.

The second independent variable is the effect sizes of the DCM measurement model parameters. The effect size is conceptually similar to the item discrimination parameter in item response theory models. It indicates the magnitude of the difference in the probability of a success response between those students classified by mastered and students classified as non-mastered. The high effect size is defined as those with a .30 higher probability for a success response along with the mastery of one additional measured attribute. The low effect size is defined as those with a .20 increase in the probability for a success response, along with the mastery of one additional measured attribute. With high effect size, classification may become more distinguishable, leading to a lower probability of label switching.

The third independent variable is the q-matrix specification. Two conditions were included: balanced and empirical, as shown in Table 4. The balanced q-matrix is specified as three times of the seven unique q-matrix patterns, with an item measuring either one attribute, two attributes, or three attributes. In contrast, the empirical q-matrix is a selected matrix from the empirical DTMR item parameters from Bradshaw, Izsák, Templin, and Jacobson (2014, p. 9). The DTMR data has four attributes and 28 items.

In order to generate the empirical q-matrix, three attributes are randomly selected from the four, without replacement, as the first step. The next step is to create the specific item pool by excluding items that measure none of the three selected attributes. Next, 21 items are randomly sampled from the item pool, with replacement. Finally, the identification of attributes is checked by multiplying the transpose of the empirical q-matrix with itself. The basic rule is that each attribute has to be measured individually by at least one item, so that it can be separated from other attributes.

All conditions of the simulation study included three attributes, 21 items, and 5000 examinees. Three has been chosen as the number of attributes in order to maintain a balance between simplicity the complexity required to mimic reality. With three attributes, there are seven ($2^3 - 1$) unique q-matrix entries, with each item measuring one, two or three attributes. With three replications of seven unique item loading patterns, 21 items are generated. The number of examinees has to be large enough for estimation stability. The number of replications of conditions has been selected in order to balance the generalizability of results versus the time required for estimation.

For each effect size and q-matrix specification, 1000 replications were generated using R (R Development Core Team, 2016; version 3.2.0), respectively. Thus 4000 unique datasets were generated using a saturated log linear structural model and a log linear cognitive diagnostic measurement model. For each unique dataset, *Mplus 7.3* (Muthen & Muthen, 1998-2016) was used to estimate parameters two times: once by a model with constraints and once by a model without constraints via. Model specifications were the same for both models, except for the constraints. There were 8000 estimations in total.

Table 5 provides a summary of values of the parameters for the structural model and Table 6 for the measurement models. All parameters were simulated from a uniform distribution in order to make the conclusion generalizable. The uniform distribution was used for all parameters, as its simplicity serves adequately for the purpose, namely generating the model parameters randomly within a certain reasonable range.

Chapter 4: Results

In total, for both the balanced q-matrix condition and the empirical q-matrix condition, there were 4000 estimations via Mplus respectively; 1000 from each of the four sub-conditions, including high effect size without constraints, high effect size with constraints, low effect size without constraints, and low effect size with constraints. In order to compare the effects of adding model constraints, each true dataset was estimated by both the model without constraints and the model with constraints. The same analytic procedures apply to both the balanced q-matrix condition and the empirical q-matrix condition. The main results for the balanced q-matrix condition and empirical q-matrix are summarized in Table 7, and showed in Figure 1 and Figure 2. Tables after Table 7 contain results regarding the parameter recovery information of the simulation study.

Balanced Q-Matrix

First, the model convergence was checked. For the model without constraints, both the high and the low effect size conditions have full convergence for their corresponding 1000 estimations. However, for the model with constraints, the convergence rate is lower. For the high effect size condition, 89.5 percent converged; whereas only 42.4 percent converged for the low effect size condition.

The second step was to identify the valid replications, which was defined as estimates from both the model without constraints and the model with constraints converge. Because all estimates using the model without constraints converge, the rate of valid replication is the same as the rate of convergence for estimation using the model with constraints. Only valid replications were used for subsequent analyses.

The third step was to check whether the model without constraints converged at the same mode as with the model with constraints. This was determined by whether the two models had the same log likelihood for the same data set. Results showed that for the high effect size condition 67.6 percent (605 out of 895) of the valid replications converged at the same mode for the model without and the model with constraints. However, only 6.1 percent (26 out of 424) of the valid replications converged at the same mode for the low effect size condition.

Furthermore, label switching was checked among the valid and same mode replications. Given the same log likelihood from the model without constraints and the model with constraints, the two models are regarded as converging at the same mode in the sample space. The models are indicated as having the same parameters if there is no label switching. In contrast, when there are differences in the models' parameters, the replication is flagged as having label switching. Structural model parameters from the two models are used for this index.

Results showed that for the high effect size condition, 86.1 percent (521 out of 605) of the valid and same mode replications had exactly the same structural model parameters. As a result, 13.9 percent were regarded as having had label switching. For the low effect size condition, 73.1 percent (19 out of 26) replications had exactly the same structural model parameters between the model without and the model with constraints, given the same log likelihood. As a result, 26.9 percent were regarded as having had label switching.

Empirical Q-Matrix

The first step was to check model convergence, which was indicated by whether a calibration provided output for the person classification parameters. For the model without constraints, both the high and the low effect size conditions have full convergence for their corresponding 1000 estimations. For the model with constraints, the convergence rate was slightly lower. For the high effect size condition, 990 replications converged; whereas 931 replications converged for the low effect size condition.

The second step was to identify the valid replications, which was defined as when calibrations from both the model without constraints and the model with constraints had converged. Because all estimates using the model without constraints converge, the rate of valid replication is the same as the rate of convergence for estimation using the model with constraints. Only valid replications were used for subsequent analyses.

The third step was to check whether the model without constraints converged at the same mode as the model with constraints. This was indicated by whether the two models had the same log likelihood for the same data set. Results showed that, for the high effect size condition, 937 out of the 990 valid replications had the same log likelihood for the model without constraints and the model with constraints. For the low effect size condition, 369 out of the 931 valid replications converged at the same mode.

As the further step, label switching was checked among the valid and same mode replications. Given the same log likelihood from the model without constraints and the model with constraints, the two models were regarded to converge at the same mode in the sample space. If no label switching occurred the models should have the same parameters. In contrast, if there were differences in their parameters the replication was flagged as having label switching. Structural model parameters from the two models were used for this index.

Results showed that for the high effect size condition, 81.4 percent (763 out of 937) of the valid and same mode replications had exactly the same structural model parameters. As a result, 18.6 percent were regarded as having had label switching. For the low effect size condition, 60 percent (221 out of 369) replications had exactly the same structural model parameters between the model without and the model with constraints, given the same log likelihood. As a result, 40 percent were regarded as having had label switching.

Parameter Recovery

For both the balanced q-matrix condition and the empirical q-matrix condition, the parameter recovery was evaluated for the structural model, the measurement model, and the person latent class classification respectively. The bias and the root mean squared error (RMSE) were examined for the converged replications across conditions.

For the structural model, the estimated log probability (i.e., the “means” from the Mplus output) was compared with its corresponding true value, which was transformed from the true latent class probability based on Equation 9. By default, Mplus constrained the last latent class log probability to zero, as a consideration of the model identification. In total, seven structural model parameters were compared between their estimated values and their true values. For the measurement model, the parameters for the balanced q-matrix condition included 21 intercepts, 36 main effects, and 21 interactions, whereas there were 21 intercepts, 23 main effects, and 2 interactions for the empirical q-matrix condition. Given the large number of parameters involved, the type of the parameter (i.e., intercepts, main effects, and interactions) was selected as the analytic unit, instead of the individual parameter for the measurement model. For the person classification accuracy, the estimated latent class was compared with the true latent class across 5000 examinees, using Cohen’s Kappa. Results from a single replication were aggregated across all converged replications within the same condition, with their standard errors reported. Results were summarized in Table 8 to Table 10.

For the structural model, as shown in Table 8, the estimated parameters recovered better for the high effect size condition than the low effect size condition. The balanced q-matrix condition had better parameter recovery than the empirical q-matrix condition. Both the mean and standard error of the RMSEs were huge for the empirical q-matrix condition, due to a few estimation outliers.

For the measurement model, as shown in Table 9, the estimated intercepts recovered the best for both q-matrix conditions, whereas the estimated interactions recovered the worst. For both conditions, the RMSE index were huge, except for the intercept parameters in the model with constraints condition.

For the person latent class classification, as shown in Table 10, the high effect size condition had much higher accuracy than the low effect size condition, for both q-matrix conditions. The classification accuracy was relatively higher for the balanced q-matrix condition than the empirical q-matrix condition. The classification accuracy was slightly higher for the model with constraints condition than the model without constraints condition.

In addition, the parameter recovery was compared between the non-label switching replications and those label switching replications. The non-label switching replications were identified when the model without constraints and the model with constraints had exactly structural model parameters, given the two models converged at the same mode (i.e., the same log likelihood). The label switching replications were identified when the two models had different structural model parameters, given the same mode. Replications that were neither converged nor shared the same mode were excluded from this analysis. Results were summarized from Table 11 to Table 14.

For the structural model, as shown in Table 11, the estimated parameters recovered better in the non-label switching replications than those label switching replications, with much smaller bias and RMSE. The label switching replications had more extreme estimated values, indicated by big RMSE values and standard errors.

For the measurement model, as shown in Table 12 for the non-label switching replications and Table 13 for the label switching replications, the non-label switching replications generally had better parameter recovery than the label switching replications. The estimated intercepts recovered well for both groups, whereas the estimated interactions recovered poorly for both.

For the person latent class classification, as shown in Table 14, both the non-label switching and the label switching replications showed similar pattern with the combined analysis mentioned previously. The high effect size condition outperformed the low effect size condition. The balanced q-matrix outperformed the empirical q-matrix condition. The classification accuracy were slightly higher and more stable in the non-label switching replications than in the label switching replications.

Chapter 5: Discussion

One distinct difference in the results between the balanced q-matrix condition and the empirical q-matrix condition was the frequency of finding local optimal solutions, when the estimation from the model without constraints had different log likelihoods from the model with constraints. For the balanced q-matrix condition 67.6 percent of valid replications converged at the same solution as for the high effect size condition, whereas only 6.1 percent converged for the low effect size condition. However, for the empirical q-matrix, the percentage was 94.6 percent for the high effect size condition and 39.6 percent for the low effect size condition.

This finding should raise concern regarding the instability in latent class analysis due to local optimal solutions. For both q-matrix conditions, the high effect size condition suffers less from the local optimal solutions than the low effect size condition. This indicates that with higher discriminating items the calibration is more likely to identify the same solution in the sample space, either without constraints or with constraints.

This finding should raise concern when using models without constraints in latent class analysis, due to the non-negligible instability of estimation. Adding constraints to a model, by definition, improves the stability of results by reducing the randomness in the search of the sample space. However, it does not guarantee a global optimal solution, even though adding constraints can stabilize the results over replications. Different start values may be recommended to test for a global optimal solution.

In addition, the estimation for the empirical q-matrix condition was more stable in both high and low effect size conditions than the balanced q-matrix condition. One reason that might account for the relative estimation stability was the relative simplicity of the empirical q-matrix, including only two two-way interactions. The balanced q-matrix had three three-way interactions, and 18 two-way interactions. The higher complexity in the q-matrix and the more unknown parameters, the more difficult for an estimation reached an optimal solution in the sample space. This indicated a simpler q-matrix design was preferred to attain higher estimation stability.

Another noteworthy finding was regarding the frequency in label switching. In the balanced q-matrix condition, for the high effect size condition, 13.9 percent of the valid replications that had the same log likelihood indicated label switching, whereas 26.9 percent of those replications were identified as label switching for the low effect size condition. In the empirical q-matrix condition, 18.6 percent of the valid replications that had the same log likelihood were identified as label switching for the high effect size condition, whereas 40 percent of those replications were flagged as showing label switching for the low effect size.

This finding should raise concern when using models without constraints in latent class analysis, due to the non-negligible high frequency of label switching, especially with low quality items. Even when the model without and the model with constraints reached the same sample space (i.e., the same log likelihood), there were occasions in which they provided different estimated parameters, indicating label switching. In such cases, the interpretation of results was confusing. By adding constraints to the model, the meaning of each latent class was predefined to reduce confusion in the result interpretation. Although adding model constraints resulted in a lower convergence rate, it provided a more accurate interpretation.

Based on the parameter recovery results, the general patterns included 1) the high effect size condition outperformed the low effect size condition, with smaller mean and standard errors of both the bias and the RMSE across replications; 2) the model with constraints outperformed the model without constraints; 3) the intercept estimation outperformed the main effect and interaction estimation in the measurement model; 4) the non-label switching replications outperformed the label switching replications in parameter recovery.

However, there were a proportion of extremely high values of both the mean and the standard errors of the RMSE across conditions, especially in the measurement model and in the empirical q-matrix condition. One potential source was the extreme estimated values in some outputs provided by Mplus when the search in the sample space was void. For example, in one replication, the estimated parameter was 6303.11, whereas the true parameter was 3.58.

Furthermore, the generalizability of the findings of the current paper is limited by the software used. The only commonly used estimation software that allowed adding the necessary constraints was Mplus. It would be interesting to compare the prevalence of label switching using other software. The frequency of label switching, as well as the convergence rate, occurrence of local optimal solutions, might differ across software.

Another potential future research direction is to compare other structural models in DCMs, such as tetrachoric correlation and Bayesian Networks. The current research uses only the saturated structural model, which is another limitation of the study that calls for future research. Other structural models may have implicit constraints on the model specification that may systematically avoid label switching because of their implicit model constraints, such as the tetrachoric correlation structural model that specifies thresholds. The prior information specified in the Bayesian Nets may have added constraints to the model as well.

References

Bradshaw, L., Izsák, A., Templin, J., & Jacobson, E. (2014). Diagnosing Teachers' Understandings of Rational Numbers: Building a Multidimensional Test within the Diagnostic Classification Framework. *Educational measurement: Issues and practice*, *33*(1), 2-14.

Celeux, G., Hurn, M., & Robert, C. P. (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, *95*(451), 957-970.

Goodman, L. A. (1974a). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61(2), 215-231.

Goodman, L. A. (1974b). The analysis of systems of qualitative variables when some of the variables are unobservable. Part IA modified latent structure approach. *American Journal of Sociology*, 1179-1259.

Grün, B., & Leisch, F. (2009). Dealing with label switching in mixture models under genuine multimodality. *Journal of Multivariate Analysis*, *100*(5), 851-861.

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*(2), 191-210.

Jasra, A., Holmes, C. C., & Stephens, D. A. (2005). Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science*, 50-67.

Kaplan, D. (Ed.). (2004). *The Sage handbook of quantitative methodology for the social sciences*. Sage Publications.

Lazarsfeld, P. F., Henry, N. W., & Anderson, T. W. (1968). *Latent structure analysis*. Boston: Houghton Mifflin.

Macready, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational and Behavioral Statistics*, *2*(2), 99-120.

Muthen, L. K., & Muthen, B. O., (1998-2015). *Mplus 7.3.*

R Development Core Team, (2016). *R 3.2.0.*

Redner, R. A., & Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, *26*(2), 195-239.

Richardson, S., & Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society. Series B (Methodological)*, 731-792.

Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement*, *6*(4), 219-262.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement*: *Theory, methods and applications.* New York, NY: The Guilford Publication Inc.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *62*(4), 795-809.

Templin, J., & Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *Journal of Classification*, *30*(2), 251-275.

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, *11*(3), 287.

Yao, W. (2012). Model based labeling for mixture models. *Statistics and Computing*, *22*(2), 337-347.

Yao, W. (2013). A Simple Solution to Bayesian Mixture Labeling. *Communications in Statistics-Simulation and Computation*, *42*(4), 800-813.

Table 1

An Example of a Q-matrix for Three Attributes

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7
Attribute 1	0	0	0	1	1	1	1
Attribute 2	0	1	1	0	0	1	1
Attribute 3	1	0	1	0	1	0	1

Note: The zeros mean an item does not measure that attribute, whereas the ones mean an item measures that attribute.

Table 2

An Example of Attribute Profiles and Probabilities for Three Attributes

	class	class	class	class	class	class	class	class	
	1	2	3	4	5	6	7	8	*Marginal*
Attribute 1	0	0	0	0	1	1	1	1	P_{α_1}
Attribute 2	0	0	1	1	0	0	1	1	P_{α_2}
Attribute 3	0	1	0	1	0	1	0	1	P_{α_3}
log probability	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6	μ_7	μ_8	$\sum \mu_c$
probability	ν_1	ν_2	ν_3	ν_4	ν_5	ν_6	ν_7	ν_8	1

Note: The zeros mean a non-mastery status of an attribute, whereas the ones refer to a mastery status of that attribute. For example, class 1 [000] refers to the group who master none of three attributes, whereas class 8 [111] refers to the group who master all three attributes; ν_c indicates the marginal probability for latent class c, whereas $\mu_c$is the log transformation of the marginal probability for class c.

Table 3

An Example of Label Switching with Macready and Dayton (1977) Data

				Replication II
Components	**Parameters**		**Replication I**	**(After label switching)**
Model fit	Log likelihood		-331.76	-331.76
Structural	Probability	Mastery	.59	.41
		Non-mastery	.41	.59
Measurement	Intercepts	Item 1	-1.33	1.12
		Item 2	-2.61	1.27
		Item 3	-3.99	-.27
		Item 4	-2.89	.88
	Main effects	Item 1	2.45	-2.45
		Item 2	3.89	-3.89
		Item 3	3.72	-3.72
		Item 4	3.78	-3.78

Table 4

The Balanced and the Empirical Q-Matrix Specification

					E_Q	
	B_Q			Attribute 1	Attribute 2	Attribute 3
Attribute 1	Attribute 2	Attribute 3	**Item #**	(PI)	(MC)	(APP)
0	0	1	1 (5)	0	1	0
0	1	0	2 (10b)	0	1	0
0	1	1	3 (2)	0	0	1
1	0	0	4 (5)	0	1	0
1	0	1	5 (18)	1	0	0
1	1	0	6 (13)	1	1	0
1	1	1	7 (19)	0	0	1
0	0	1	8 (10a)	0	1	0
0	1	0	9 (10c)	0	1	0
0	1	1	10 (3)	1	0	0
1	0	0	11 (17)	1	0	0
1	0	1	12 (11)	0	1	0
1	1	0	13 (11)	0	1	0
1	1	1	14 (19)	0	0	1
0	0	1	15 (8c)	0	0	1
0	1	0	16 (15b)	1	1	0

0	1	1	17 (19)	0	0	1
1	0	0	18 (3)	1	0	0
1	0	1	19 (8b)	0	0	1
1	1	0	20 (22)	1	0	0
1	1	1	21 (22)	1	0	0

Note: B_Q refers to the balanced q-matrix condition, and the E_Q refers to the empirical q-matrix condition. Information in parentheses refers to the corresponding attributes and item numbers in Table 1 from Bradshaw, Izsák, Templin, and Jacobson (2014, p. 9).

Table 5

The Structural Model Parameter Specification in the Simulation Study

Structural Models	**Parameters**	**Symbols**	**Distribution**
Saturated log linear	Main effect	$(\gamma_{i,1,a})$	$U(-1.5, 1.5)$
	Two-way interaction	$(\gamma_{i,2,(a,a')})$	$U(-1.5, 1.5)$
	Three-way interaction	$(\gamma_{i,3,(1,2,3)})$	$U(-1.5, 1.5)$

Note: U(a,b) refers to a uniform distribution, ranging from a to b.

Table 6

The Measurement Model Parameter Specification in the Simulation Study

Effect		**Items Measuring**		
Size	**Parameter**	**One Attribute**	**Two Attributes**	**Three Attributes**
High	Intercept	U(-1.62, .38)	U(-2.39, -.39)	U(-3.94, -1.94)
	Main effect	U(.49, 1.99)	U(.64, 2.14)	U(1.58, 3.08)
	Two-way interaction	NA	U(-.50, .50)	U(-1.59, -.59)
	Three-way interaction	NA	NA	U(-.65, .35)
Low	Intercept	U(-1.41, .59)	U(-1.85, 0.15)	U(-2.39, -.39)
	Main effect	U(.06, 1.56)	U(0.10, 1.60)	U(0.23, 1.73)
	Two-way interaction	NA	U(-0.50, 0.50)	U(-0.67, .33)
	Three-way interaction	NA	NA	U(-1.14, -.14)

Note: NA indicates inapplicable; U(a, b) means a uniform distribution, ranging from a to b.

Table 7

Main Results for the Balanced Q-Matrix and Empirical Q-Matrix

	Effect Size	**NC_ convergence**	**C_ convergence**	**Valid Replications**	**Valid & ΔLL=0**	**Valid & ΔLL=0 & Δμ=0**
B_Q	High	1000	895	895	605	521
	Low	1000	424	424	26	19
E_Q	High	1000	990	990	937	763
	Low	1000	931	931	369	221

Note: 1000 replications for the high and low effect size respectively. B_Q refers to the balanced q-matrix condition, and E_Q refers to the empirical q-matrix condition; NC refers to the model without constraints, C refers to the model with constraints, Δ refers to the difference, LL refers to the log likelihood, μ refers to the structural model parameters.

Table 8

Parameter Recovery Results for the Structural Model Parameters

		B_Q					E_Q				
			Bias		RMSE			Bias		RMSE	
		#	Mean	SE	Mean	SE	#	Mean	SE	Mean	SE
H	NC	1000	-0.17	0.73	0.75	3.11	999	0.52	23.21	1164.41	21091.04
	C	895	-0.18	0.61	0.53	1.57	990	6.75	200.40	40444.52	1261173.14
L	NC	1000	-0.32	8.51	496.05	15578.06	996	-3.06	64.43	16701.20	211485.07
	C	424	-0.11	1.01	1.43	2.34	931	11.56	375.30	142519.49	4285605.13

Note: B_Q refers to the balanced q-matrix condition, and E_Q refers to the empirical q-matrix condition. H refers to the high effect size condition, and L refers to the low effect size condition. NC refers to the model without constraints condition, and C refers to the model with constraints condition. The # refers to the number of replications aggregated over to calculate the mean and standard errors of the bias and the RMSE. It is worth mentioning that the # could vary across the type of parameters within the same condition, because a converging replication may not report a specific parameter, such as an interaction term. SE refers to the standard error.

Table 9

Parameter Recovery Results for the Measurement Model Parameters

		B_Q					E_Q				
			Bias		RMSE			Bias		RMSE	
		#	Mean	SE	Mean	SE	#	Mean	SE	Mean	SE
Intercept		(21)					(21)				
H	NC	999	-0.74	5.21	518.48	6725.38	999	-0.06	1.39	41.75	1313.87
	C	895	-0.05	0.21	0.57	1.59	990	-0.01	0.06	0.04	0.27
L	NC	999	-0.51	8.46	1494.99	42220.47	999	-0.99	16.01	5345.68	128873.05
	C	424	0.02	0.14	0.15	0.98	931	-0.03	0.12	0.18	0.92
Main Effect		(36)					(23)				
H	NC	999	7.54	92.83	288837.82	5836478.32	1000	1.07	30.80	10908.99	342502.42
	C	895	0.38	3.32	390.77	8175.64	990	0.04	0.71	11.63	347.72
L	NC	999	7.04	74.32	176399.72	3315690.53	1000	3.02	45.46	24136.62	518055.36
	C	424	2.53	28.70	29372.27	420297.33	930	8.24	134.64	416734.47	10509019.84

Interaction (21)							(2)				
H	NC	990	0.24	34.39	35538.25	514947.15	999	-0.30	15.32	466.60	14549.52
	C	895	3.33	32.35	36492.59	361819.41	990	1.28	19.07	720.64	15044.32
L	NC	978	10.07	179.36	731297.30	11360077.85	999	6.53	504.97	509319.52	14468828.91
	C	422	15.33	107.48	266392.31	2627872.29	927	24.02	900.10	1614690.46	47523475.05

Note: B_Q refers to the balanced q-matrix condition, and E_Q refers to the empirical q-matrix condition. H refers to the high effect size condition, and L refers to the low effect size condition. NC refers to the model without constraints condition, and C refers to the model with constraints condition. The # refers to the number of replications aggregated over to calculate the mean and standard errors of the bias and the RMSE. SE refers to the standard error.

Table 10

Parameter Recovery Results for the Person Latent Class Classifications

			B_Q			E_Q	
			Kappa			Kappa	
		#	Mean	SE	#	Mean	SE
H	NC	1000	0.55	0.09	1000	0.46	0.05
	C	895	0.56	0.07	990	0.46	0.05
L	NC	1000	0.32	0.10	1000	0.26	0.08
	C	424	0.34	0.08	931	0.27	0.07

Note: B_Q refers to the balanced q-matrix condition, and E_Q refers to the empirical q-matrix condition. H refers to the high effect size condition, and L refers to the low effect size condition. NC refers to the model without constraints condition, and C refers to the model with constraints condition. The # refers to the number of replications aggregated over to calculate the mean and standard errors of the bias and the RMSE. SE refers to the standard error.

Table 11

Parameter Recovery Results for the Structural Model Parameters between the Non-Label Switching and the Label Switching Replications

		B_Q					E_Q				
			Bias		RMSE			Bias		RMSE	
		#	Mean	SE	Mean	SE	#	Mean	SE	Mean	SE
Non Label Switching											
H	NC	521	-0.19	0.50	0.37	1.06	763	-0.07	0.42	0.63	3.35
	C	521	-0.19	0.50	0.37	1.06	763	-0.07	0.44	0.77	4.50
L	NC	19	0.00	0.44	0.43	0.59	221	-0.29	0.90	3.65	9.64
	C	19	0.00	0.44	0.43	0.59	221	-0.51	3.20	60.99	840.90
Label Switching											
H	NC	84	-0.29	0.67	0.61	1.58	174	2.26	8.08	79.15	430.97
	C	84	-0.29	0.67	0.61	1.58	174	39.00	477.48	228200.50	3008194.89
L	NC	7	-0.59	0.64	0.97	1.17	148	1.22	12.27	528.72	5187.44
	C	7	-0.59	0.64	0.96	1.17	148	1.32	5.39	40.14	89.02

Note: B_Q refers to the balanced q-matrix condition, and E_Q refers to the empirical q-matrix condition. H refers to the high effect size condition, and L refers to the low effect size condition. NC refers to the model without constraints condition, and C refers to the model with constraints condition. The # refers to the number of replications aggregated over to calculate the mean and standard errors of the bias and the RMSE. SE refers to the standard error.

Table 12

Parameter Recovery Results for the Measurement Model Parameters for the Non-Label Switching Replications

			B_Q					E_Q			
			Bias		RMSE			Bias		RMSE	
		#	Mean	SE	Mean	SE	#	Mean	SE	Mean	SE
Intercept (21)							(21)				
H	NC	521	-0.03	0.24	0.99	10.46	763	0.00	0.04	0.02	0.04
	C	521	-0.02	0.13	0.26	1.22	763	0.00	0.04	0.02	0.04
L	NC	19	0.03	0.11	0.04	0.04	221	-0.01	0.05	0.02	0.03
	C	19	0.03	0.11	0.04	0.04	221	-0.01	0.05	0.02	0.03
Main Effect (36)							(23)				
H	NC	521	0.10	0.40	1.87	18.56	763	0.01	0.06	0.05	0.31
	C	521	0.07	0.21	0.61	3.21	763	0.01	0.06	0.05	0.35
L	NC	19	0.06	0.08	0.11	0.06	221	0.05	0.14	0.36	1.98
	C	19	0.06	0.08	0.11	0.06	221	0.69	9.52	2081.18	30935.13
Interaction (21)							(2)				
H	NC	521	0.96	19.11	7655.50	172922.50	763	0.16	1.28	2.95	18.50
	C	521	0.26	2.50	1880.48	23984.76	763	0.43	7.26	105.77	2774.38

L	NC	19	0.07	0.37	14.25	20.82	221	0.31	4.77	44.92	550.70
	C	19	0.02	0.22	12.60	17.55	221	-6.89	109.66	23996.60	356012.22

Note: B_Q refers to the balanced q-matrix condition, and E_Q refers to the empirical q-matrix condition. H refers to the high effect size condition, and L refers to the low effect size condition. NC refers to the model without constraints condition, and C refers to the model with constraints condition. The # refers to the number of replications aggregated over to calculate the mean and standard errors of the bias and the RMSE. SE refers to the standard error.

Table 13

Parameter Recovery Results for the Measurement Model Parameters for the Label Switching Replications

		B_Q					E_Q				
			Bias		RMSE			Bias		RMSE	
		#	Mean	SE	Mean	SE	#	Mean	SE	Mean	SE
Intercept (21)							(21)				
H	NC	84	-0.23	1.37	38.75	343.23	174	-0.01	0.06	0.06	0.43
	C	84	-0.05	0.16	0.42	1.02	174	-0.01	0.05	0.03	0.13
L	NC	7	-0.03	0.06	0.02	0.01	148	-0.01	0.05	0.03	0.04
	C	7	-0.03	0.06	0.02	0.01	148	-0.01	0.05	0.03	0.04
Main Effect (36)							(23)				
H	NC	84	0.44	2.39	68.10	600.09	174	0.03	0.10	0.11	0.78
	C	84	0.12	0.25	0.93	2.29	174	0.02	0.08	0.07	0.23
L	NC	7	0.07	0.04	0.09	0.03	148	0.09	0.22	0.78	3.52
	C	7	0.07	0.04	0.09	0.03	148	25.20	305.43	214270.81	26067064.11
Interaction (21)							(2)				

H	NC	84	-0.28	2.92	171.97	1371.33	174	0.53	2.38	10.07	34.19
	C	84	5.09	45.69	43583.81	398734.49	174	2.39	22.38	1011.01	13085.61
L	NC	7	0.08	0.37	39.99	58.82	148	0.08	3.68	19.07	62.38
	C	7	0.04	0.30	29.60	51.99	147	0.82	9.12	159.27	1353.55

Note: B_Q refers to the balanced q-matrix condition, and E_Q refers to the empirical q-matrix condition. H refers to the high effect size condition, and L refers to the low effect size condition. NC refers to the model without constraints condition, and C refers to the model with constraints condition. The # refers to the number of replications aggregated over to calculate the mean and standard errors of the bias and the RMSE. SE refers to the standard error.

Table 14

Parameter Recovery Results for the Person Latent Class Classifications between the Non-Label Switching and the Label Switching Replications

			B_Q Kappa Mean			E_Q Kappa Mean	
		#		SE	#		SE
Non Label Switching							
H	NC	521	0.58	0.05	763	0.47	0.04
	C	521	0.58	0.05	763	0.47	0.04
L	NC	19	0.36	0.08	221	0.30	0.05
	C	19	0.36	0.08	221	0.30	0.05
Label Switching							
H	NC	84	0.56	0.06	174	0.46	0.05
	C	84	0.56	0.06	174	0.46	0.05
L	NC	7	0.38	0.05	148	0.28	0.07
	C	7	0.38	0.05	148	0.28	0.07

Note: B_Q refers to the balanced q-matrix condition, and E_Q refers to the empirical q-matrix condition. H refers to the high effect size condition, and L refers to the low effect size condition. NC refers to the model without constraints condition, and C refers to the model with constraints condition. The # refers to the number of replications aggregated over to calculate the mean and standard errors of the bias and the RMSE. SE refers to the standard error.

Figure 1

Main Results for the Balanced Q-Matrix Condition

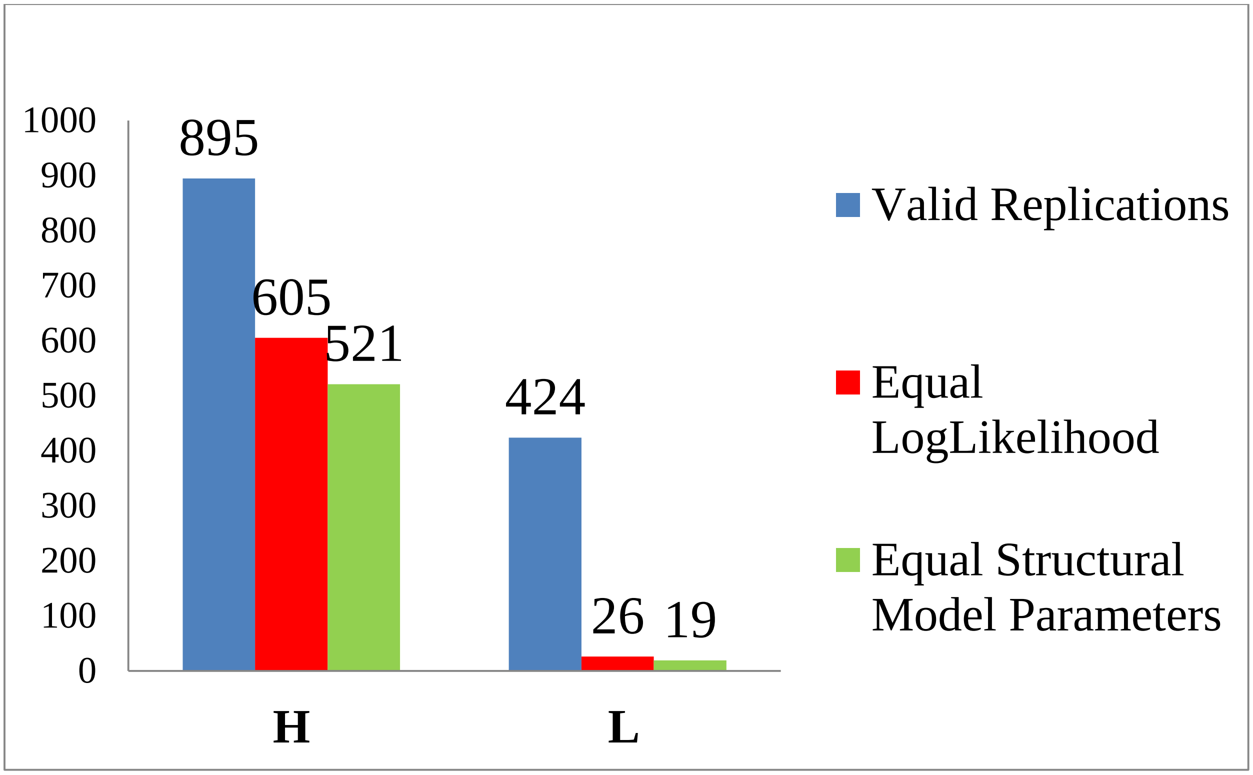
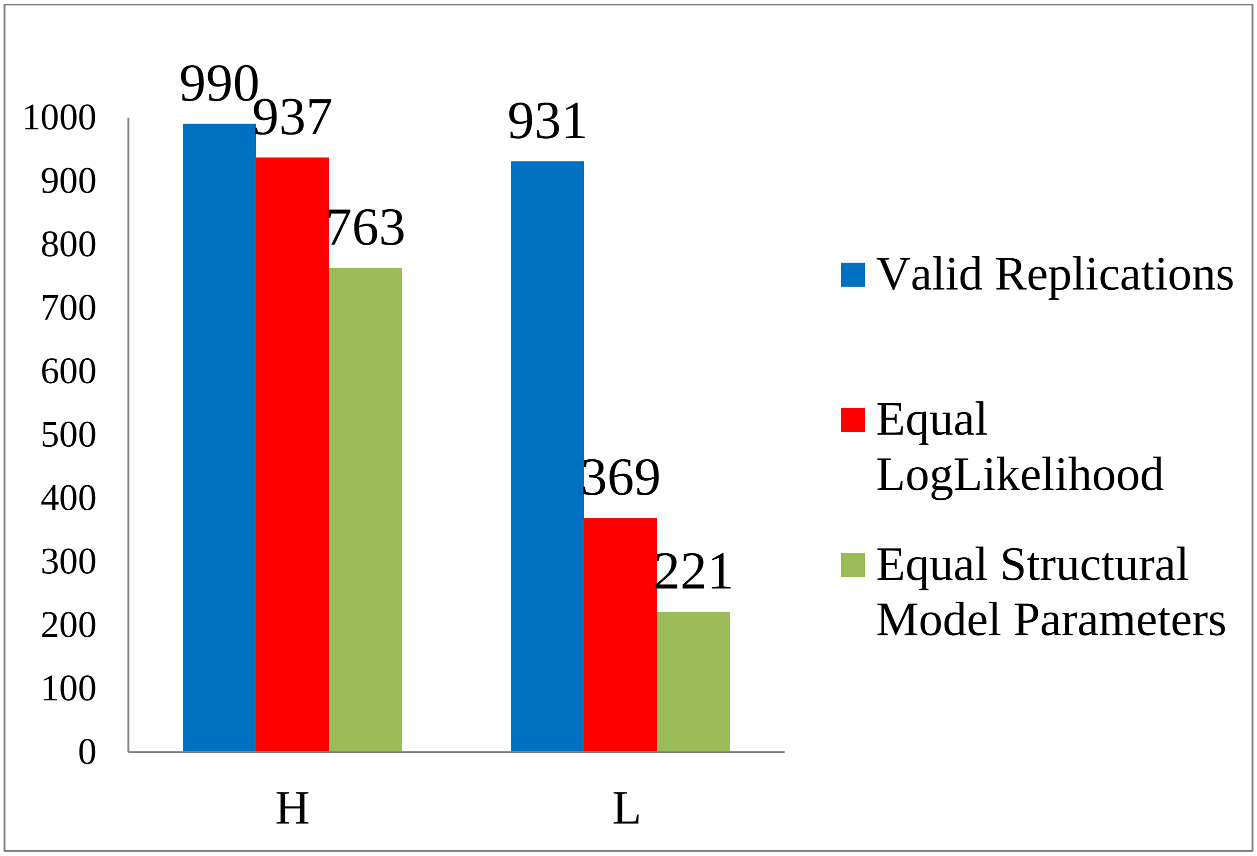


Figure 2

Main Results for the Empirical Q-Matrix Condition


H
L
Valid Replications
Equal LogLikelihood
Equal Structural Model Parameters