COGNITIVE DIAGNOSIS MODELING AND APPLICATIONS TO ASSESSING LEARNING

BY

SUSU ZHANG

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Psychology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2018

Urbana, Illinois

Doctoral Committee:

Professor Hua-Hua Chang, Chair
Professor Carolyn J. Anderson
Professor Steven A. Culpepper
Professor Jeffery A. Douglas
Professor Jinming Zhang

Abstract

Chapter 1: Cognitive diagnosis models (CDMs) are restricted latent class models designed to assess test takers' mastery on a set of skills or attributes. With a wide range of applications in education and in psychopathology, various CDMs have been proposed and fitted to response data from different scenarios. Recently, Xu (2017) derived sufficient conditions for identifying model parameters of a restricted latent class model, which generalizes many existing CDMs. We propose a Bayesian estimation algorithm for this restricted latent class model. The model is applied to the Examination for the Certificate of Proficiency in English language assessment data (e.g. Henson & Templin, 2007).

Chapter 2: There has been a growing interest in measuring students' growth over time. CDMs were traditionally used to measure students' skill mastery at a static time point, but recently, they have been used in many longitudinal models to track students' changes in skill acquisition over time. In this chapter, we propose a longitudinal learning model, where different kinds of skill hierarchies were considered, and the reduced-reparameterized unified model (r-RUM) or the noisty input, deterministic-"and"-gate (NIDA) model is used to measure students' skill mastery at each time point. This model is fitted to the Spatial Rotation data set (e.g., S. Wang, Yang, Culpepper, & Douglas, 2016), and different models were compared using Bayesian model comparison methods.

Chapter 3: The increased popularity of computer-based testing has enabled researchers to collect various types of process data, including response times. Extensive research has been conducted on the joint modeling of response accuracy and response times. Recent research on CDMs begins to explore the relationship between speed and accuracy to understand

students' fluency of applying the mastered skills, in addition to mastery information, in a learning environment. In this chapter, we propose a mixture hidden Markov Diagnostic Classification Model framework for learning with response times and response accuracy. Such a model accounts for the heterogeneities in learning styles among students by modeling the different learning and response behaviors among subgroups. The proposed model is evaluated through a simulation study in terms of parameter recovery.

Chapter 4: We introduce an R package, `hmcdm`, that can be used to fit several longitudinal models for learning under the cognitive diagnosis framework. The package allows users to simulate item responses (and response times if applicable) under several learning models, to fit the models using Markov Chain Monte Carlo (MCMC) methods, to compute point estimates of parameters based on the MCMC samples, and to evaluate and compare different models using Deviance Information Criterion and posterior predictive probabilities.

To my family.

Acknowledgments

I want to first express my deep gratitude for Dr. Hua-Hua Chang, my advisor. My interest in psychometrics and learning were sparked by you. And for my entire time here, you never stopped caring about me, encouraging me, and advising me. I have lost count of how much you have taught me, how many opportunities you have brought me, and how many conferences I have attended with your support. I will strive to become a wonderful person like you.

My growth is highly indebted to my professors. A special thanks to Dr. Steven Culpepper and Dr. Jeffrey Douglas, for their generosity in sharing research ideas, for their tremendous amount of guidance on my research, and for their kindness to always help me and introduce me to various opportunities. I have learned so much from their classes and from working with them, and the first two chapters of my dissertation could not have been completed without them. I would also like to thank Dr. Carolyn Anderson and Dr. Jinming Zhang, for sharing the burden as committee members, for teaching me so many things, for always being available to answer my questions, and for all the valuable suggestions they have provided on my research. In addition, I would like to show my appreciations to Dr. Hans-Friedrich Köhn, who was not on my committee, but has enlightened me, encouraged me, and helped me just as much. Each one of you have taught me an immense amount, and I will try my best to use what I have learned to contribute to the field and to help others.

I could not have made it this far without the support from my peers. Especially, I would like to thank Justin Kern and Edison Choe, who have helped me both in my studies and in my life since day 1; Shiyu Wang, from whom I learned so much about rigorous research

and without whom Chapter 3 of my dissertation would not have been done; and James Balamuta, who kept broadening my horizons on efficient and reproducible computing, and without whom the R package in Chapter 4 would look much cheaper. The path to a Ph.D. was not the easiest journey, and I was lucky to be comrades with Anqi Li, Xiao Li, Yang Du, Alvaro Cruz, Auburn Jimenez, and many others.

Lastly, I am thankful to my parents, Lei Zhang and Kun Bao, and my boyfriend, Tianhao Wu, for their unwavering support and company and for helping me grow as a person. I will always remember not to take what I have as granted, and to always do what I can to help those in need.

Contents

List of Tables . . . . . **ix**

List of Figures . . . . . **x**

Chapter 1 Bayesian Estimation of Restricted Latent Class Models **1**
1.1 Introduction 1
1.2 Xu's (2017) Restricted Latent Class Model 2
1.2.1 The RLCM 3
1.2.2 Identifiability of CDMs and Xu's Sufficient Conditions 4
1.2.3 Special Cases 5
1.2.4 Bayesian Formulation 8
1.3 Simulation Studies 10
1.3.1 True parameter generation 10
1.3.2 Parameter Estimation 11
1.3.3 Evaluation Criteria 13
1.3.4 Results 14
1.4 Application: Examination for the Certificate of Proficiency in English 23
1.5 Discussion 30

Chapter 2 Assessing Learning with the Reduced RUM **34**
2.1 Introduction 34
2.1.1 Learning Models Based on CDMs 35
2.2 Current Model 36
2.2.1 Learning Model 37
2.2.2 Measurement Models 38
2.3 Parameter Estimation 39
2.3.1 Prior distribution 39
2.3.2 Full conditional distributions 40
2.3.3 A Gibbs Sampling Algorithm 43
2.4 Application: A Spacial Reasoning Test with Learning Interventions 44
2.4.1 Spatial Rotation Data 44
2.4.2 Evaluated Models 46
2.4.3 Model convergence 47
2.4.4 Model Comparison 48
2.4.5 Observed progression of learning 52
2.5 Discussion 54

Chapter 3 Mixture Learning Model with Responses and Response Times 55
3.1 Introduction . . . 55
3.1.1 Joint Learning Model with Response Times and Response Accuracy . 57
3.1.2 Mixture Models and Mixture Hidden Markov Models . . . 59
3.2 Mixture Learning Model with Response Times and Response Accuracy . . . 61
3.2.1 Bayesian Model Formulation . . . 64
3.2.2 Bayesian Full Conditional Distribution . . . 66
3.3 Simulation Study . . . 72
3.3.1 True Parameters . . . 73
3.3.2 Parameter Estimation . . . 74
3.3.3 Evaluation Criteria . . . 75
3.3.4 Results . . . 76
3.4 Discussion . . . 79

Chapter 4 `hmcdm`: An R Package for Fitting Learning Models . . . 83
4.1 Introduction . . . 83
4.2 Availability . . . 84
4.3 Documentations . . . 84

References . . . 85

List of Tables

1.1 Attribute-wise Agreement Rate (AAR), Pattern-wise Agreement Rate (PAR), and computational Time of the RLCM Gibbs Sampler for Values of K, ρ, and N. 24

2.1 Summary of fit statistics of the six different models. M1 misfit represents the percentage of item means that were outside the 95% posterior prediction interval, M2 misfit represnets percentage of item pair-wise odds ratios outside the prediction interval, and total misfit represents the percentage of observed total scores at each time point outside the 95% posterior prediction interval. 50
2.2 Frequency (and percentage) of number of skills mastered over time. 54

3.1 Components of the mixture learning model with disengaged and engaged test takers. 64
3.2 The attribute-wise and pattern-wise agreement rates (AARs and PARs) between the true and estimated $\boldsymbol{\alpha}$. 78
3.3 The true and estimated covariance between $\boldsymbol{\theta}$ and $\boldsymbol{\tau}$ ($\boldsymbol{\Sigma}$), learning model parameters ($\boldsymbol{\lambda}$), mixing weight (ω), coefficient for speed growth (ϕ_0) in the engaged learning mode, correct response probability of learners in the disengaged mode (g^*), and the log-normal mean (μ_1) and standard deviation (σ_1) of the response time distribution in the disengaged mode. "True" stands for the true value of the parameters, "EAP" is the average of the parameter samples across iterations, and "SD" is the standard deviation of the parameter samples across iterations. 79
3.4 Root mean square errors (RMSE) and correlations between true and estimated DINA item parameters ($\mathbf{s}, \mathbf{g}$), response time model parameters ($\mathbf{a}, \boldsymbol{\gamma}$), and latent speed ($\boldsymbol{\tau}$) and learning ability ($\boldsymbol{\theta}$) of learners. 79

List of Figures

1.1 Brooks-Gelman $\hat{R}$ for the convergence of $\boldsymbol{\Theta}$ and $\boldsymbol{\pi}$ in 5 chains, under the condition of $N = 2500, K = 4, \rho = .5$. Horizontal line represents the $\hat{R} = 1.2$ cutoff that is commonly used as a threshold for parameter convergence. . . . 13

1.2 Bias of item parameter (i.e., elements of $\boldsymbol{\theta}$) estimates across repetitions when the number of attributes is $K = 3$. *Note.* The three different plotting characters item parameters for three groups based on the type of skill mastery pattern: a cross denotes parameters for the class without any attributes (i.e., $\boldsymbol{\alpha} = \mathbf{0}$); a hollow dot represents classes that do not have all required attributes (i.e., $\boldsymbol{\alpha} \succ \mathbf{0}$ and $\boldsymbol{\alpha} \not\succeq \mathbf{q}_j$); and a solid dot indicates classes that have all requisite skills (i.e., $\boldsymbol{\alpha} \succeq \mathbf{q}_j$). 15

1.3 RMSE of item parameter (i.e., elements of $\boldsymbol{\theta}$) estimates for $K = 3$. *Note.* The three different plotting characters item parameters for three groups based on the type of skill mastery pattern: a cross denotes parameters for the class without any attributes (i.e., $\boldsymbol{\alpha} = \mathbf{0}$); a hollow dot represents classes that do not have all required attributes (i.e., $\boldsymbol{\alpha} \succ \mathbf{0}$ and $\boldsymbol{\alpha} \not\succeq \mathbf{q}_j$); and a solid dot indicates classes that have all requisite skills (i.e., $\boldsymbol{\alpha} \succeq \mathbf{q}_j$). 16

1.4 Bias of item parameter (i.e., elements of $\boldsymbol{\theta}$) estimates for $K = 4$. *Note.* The three different plotting characters item parameters for three groups based on the type of skill mastery pattern: a cross denotes parameters for the class without any attributes (i.e., $\boldsymbol{\alpha} = \mathbf{0}$); a hollow dot represents classes that do not have all required attributes (i.e., $\boldsymbol{\alpha} \succ \mathbf{0}$ and $\boldsymbol{\alpha} \not\succeq \mathbf{q}_j$); and a solid dot indicates classes that have all requisite skills (i.e., $\boldsymbol{\alpha} \succeq \mathbf{q}_j$). 18

1.5 RMSE of item parameter (i.e., elements of $\boldsymbol{\theta}$) estimates for $K = 4$. *Note.* The three different plotting characters item parameters for three groups based on the type of skill mastery pattern: a cross denotes parameters for the class without any attributes (i.e., $\boldsymbol{\alpha} = \mathbf{0}$); a hollow dot represents classes that do not have all required attributes (i.e., $\boldsymbol{\alpha} \succ \mathbf{0}$ and $\boldsymbol{\alpha} \not\succeq \mathbf{q}_j$); and a solid dot indicates classes that have all requisite skills (i.e., $\boldsymbol{\alpha} \succeq \mathbf{q}_j$). 19

1.6 Bias for population membership probability ($\boldsymbol{\pi}$) estimates when $K = 3$. . . . 20

1.7 RMSE for population membership probability ($\boldsymbol{\pi}$) estimates when $K = 3$. . . 21

1.8 Bias for population membership probability ($\boldsymbol{\pi}$) estimates when $K = 4$. . . . 22

1.9 RMSE for population membership probability ($\boldsymbol{\pi}$) estimates when $K = 4$. . . 23

1.10 Posterior predictive probabilities of observed total scores for the RLCM and the rRUM. 26

1.11 Plots of RLCM empirical item response probabilities across attribute profiles $\boldsymbol{\alpha}_c$. Note. Some of the band widths were very close to 0, they are not shown on the figures. 28
1.12 Plots of rRUM empirical item response probabilities across attribute profiles $\boldsymbol{\alpha}_c$. Note. Some of the band widths were very close to 0, they are not shown on the figures. 29
1.13 Posterior predictive probabilities of the observed item pairwise odds ratios for the RLCM (top-left) and the rRUM (bottom-right). Each bubble represents an item pair, and the shaded area on each bubble represents the posterior predictive probability of the observed odds ratio for that pair of item. 31

2.1 Test Block 45
2.2 Learning Block: Type 1 45
2.3 Learning Block: Type 2 45
2.4 Progression of maximum $\hat{R}$ as chain length increases. Dashed horizontal line indicates $\hat{R} = 1.2$. 48
2.5 Posterior Predictive Probabilities (PPPs) of the item means (i.e., proportion correct). 51
2.6 Density of the posterior predictive probability for item pair-wise odds ratios. 51
2.7 Density of posterior predictive probability for total scores at different time points. 52
2.8 Relationship between posterior predictive probability of total score and observed total score at each time point. 53
2.9 Progression of mastery rate of each skill across time. 53

3.1 Maximum Brook-Gelman Proportional Scale Reduction Factor across all parameters with different chain lengths. The $x-$axis is the length of the MCMC chain, and the $y-$axis is the maximum PRSF. Dashed line represents the commonly used threshold of $\hat{R} = 1.2$ for parameter convergence. 77

Chapter 1

Bayesian Estimation of Restricted Latent Class Models

1.1 Introduction

Cognitive diagnosis models (CDMs; Rupp, Templin, & Henson, 2010) are restrictive latent class models (RLCMs) measuring test-takers' underlying attribute, or skill patterns, based on their responses to test items. CDMs could be used in the education context to identify students' strengths and weaknesses in learning, providing guidance for personalized, targeted instruction and support. In the psychopathological context CDMs could be applied to estimate the test takers' underlying pattern of symptoms, helping diagnosis to design treatments. With a broad range of practical implications, various CDMs have been proposed by previous researchers (e.g., Junker & Sijtsma, 2001; de la Torre, 2011; Henson, Templin, & Willse, 2009; von Davier, 2008) to model different relationships between responses to items and test takers' underlying attribute patterns.

For any model, identifiability needs to be established as a prerequisite to the estimation and inference of the model parameters. In fact, several studies (e.g., see Chen, Liu, Xu, & Ying, 2015; Xu & Zhang, 2016) have established identifiability conditions for the more restrictive deterministic input, noisy-"and"-gate (DINA, Junker & Sijtsma, 2001) model. More recently, Xu (2017) proposed a set of sufficient conditions to ensure identifiability of restricted latent class models (RLCMs) for dichotomous responses. Xu's identifiability constraints involve restrictions on the item parameters and the Q-matrix, and he showed that many of the previously developed CDMs are special cases of the RLCM.

We propose a Bayesian modeling framework for the RLCM and introduce a Gibbs sam-

pling algorithm that jointly estimates the item, examinee, and population class parameters based on an observed response matrix. Such an estimation algorithm for this general set of CDMs could be highly applicable for item calibrations where assumptions behind more restrictive models are not satisfied. Furthermore, by including additional restrictions to the parameters, the RLCM could be made equivalent to the more restrictive and interpretable models, and thus the proposed estimation algorithm could serve as a versatile tool in model selection, where a test developer can start from fitting the most relaxed model and gradually move on to more restricted and interpretable ones.

In this chapter, we provide a brief introduction to cognitive diagnosis models and model identifiability, introduce the general class of restricted latent class models proposed in Xu (2017), and re-state the theorems on the identifiability of the RLCM parameters. A few examples of commonly used CDMs will be demonstrated to show how they satisfy Xu's identifiability restrictions. A Bayesian model set-up for the class of RLCMs satisfying the identifiability constraints will be presented. We derive the model parameter full conditional distributions for a Gibbs sampling algorithm. The results from a simulation study conducted to evaluate the performance of the proposed estimation routine under different conditions are reported. And lastly, we conduct an empirical study using the Examination for the Certificate of Proficiency in English (ECPE) language assessment data (Henson & Templin, 2007; Templin & Hoffman, 2013) and compare model estimates between the RLCM and the reduced reparametrized unified model (rRUM, Hartz, 2002).

1.2 Xu's (2017) Restricted Latent Class Model

This section discusses Xu's (2017) RLCM. The first subsection introduces the RLCM and outlines the sufficient conditions to ensure model identifiability. The second subsection discusses how several existing CDMs are special cases of the RLCM, and the third subsection describes the Bayesian formulation and Gibbs sampler for approximating the posterior dis-

tribution.

1.2.1 The RLCM

CDMs are concerned with classifying individuals as either masters or non-masters on a set of attributes. Throughout the following discussion we denote $\{1, 2, \ldots, D\}$ by $[D]$. Let $k \in [K]$ index attributes and $i \in [N]$ index individuals. The binary attribute pattern for individual i is $\boldsymbol{\alpha}_i = [\alpha_{i1}, \ldots, \alpha_{iK}]$ where $\alpha_k = 1$ if the test-taker has mastered the kth attribute, and $\alpha_k = 0$ otherwise. CDMs assume that $\boldsymbol{\alpha}_i$ underlies performance on a collection of binary tasks/items. Specifically, let $\boldsymbol{X}_i = [X_{i1}, \ldots, X_{iJ}]$ denote individual i's observed response vector for $j \in [J]$ where X_{ij} is one for correct responses and zero otherwise. Furthermore, the probability of responding "1" on X_{ij} is a function of test taker's attribute profile and the requisite attributes required for item j. In other words, CDMs require a $J \times K$ Q matrix where $Q_{jk} = 1$ if correctly responding to item j is related to the mastery of attribute k, and $Q_{jk} = 0$ otherwise.

Under the RLCM, the model for the jth response for individual i with attribute pattern $\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_c$ is,

$$X_{ij} \,|(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_c, \boldsymbol{\Theta}) \sim Bernoulli(\theta_{\boldsymbol{\alpha}_c, j}), \tag{1.1}$$

where $\theta_{\boldsymbol{\alpha}_c, j}$ is the correct response probability to item j, given attribute pattern $\boldsymbol{\alpha}_c$. Let the matrix of item by attribute class matrix $\boldsymbol{\Theta}$ be defined as,

$$\boldsymbol{\Theta} = \begin{pmatrix} \theta_{\boldsymbol{\alpha}_1,1} & \ldots & \theta_{\boldsymbol{\alpha}_C,1} \\ \vdots & \ddots & \vdots \\ \theta_{\boldsymbol{\alpha}_1,J} & \ldots & \theta_{\boldsymbol{\alpha}_C,J} \end{pmatrix}.$$

An immediate observation is that the RLCM admits at most 2^K distinct response probabilities for items with $Q_{jk} = 1$ for all k. As discussed in the next section, some restrictions must be placed upon $\boldsymbol{\Theta}$ to ensure model identifiability.

1.2.2 Identifiability of CDMs and Xu's Sufficient Conditions

Unrestricted latent class models with binary responses are not in general identifiable (Lazarsfeld, 1950), thus for any restricted latent class model, such as CDMs, identifiability needs to be established before proceeding to model estimation and inferences. We say a vector of parameters β from a family of distributions $f(x \mid \beta), \beta \in B$ is identifiable if the parameters β that can generate probability density function $f(x \mid \beta)$ is unique. In other words, for any $\beta, \beta' \in B$, $f(x \mid \beta) = f(x \mid \beta')$ for $\forall x$ only if $\beta = \beta'$ (Xu, 2017). Without model identifiability, given any observed data, two different sets of model parameters can generate the same likelihood, thus an unique estimate for the parameters could not be obtained. It can also be shown that model identifiability is a necessary condition for estimation consistency (Gabrielsen, 1978).

Several previous studies have looked into the conditions under which specific CDMs could be identifiable. Chiu, Douglas, and Li (2009) proved that under the DINA model when the item parameters are known, having a complete Q-matrix (i.e., having at least one item measuring each attribute and that attribute only) is the sufficient and necessary condition for the identifiability of the membership probabilities for attribute classes in the population. Xu and Zhang (2016) further showed that when the item parameters are unknown for the DINA model, a more restrictive Q-matrix structure would be sufficient to guarantee the identifiability of the item and population parameters.

Xu (2017) recently extended the sufficient conditions for identifiability for the DINA model to more general restrictive latent class models. In other words, he showed that the item parameters, $\boldsymbol{\Theta}$, and the population membership probabilities of the attribute classes, $\boldsymbol{\pi}$, are identifiable if restrictions are placed on the item parameters $\boldsymbol{\Theta}$ and the Q-matrix structure (see Xu, 2017, for a proof and examples). Let $\boldsymbol{q}_j$ denote the jth row of Q and note that $\boldsymbol{\alpha} \succeq \mathbf{q}_j$ represents $\alpha_k \geq q_{jk}$, for all k, let $\boldsymbol{e}_k$ denote the unit vector with 1 on the $k-$th entry and 0 elsewhere, and let $\boldsymbol{\alpha}' \not\succeq \mathbf{q}_j$ represent $\alpha_k < q_{jk}$ for some k. Three restrictions must be in place to ensure model identifiability:

- $\theta_{\boldsymbol{q},j} = \max_{\boldsymbol{\alpha}:\boldsymbol{\alpha}\succeq\mathbf{q}_j} \theta_{\boldsymbol{\alpha},j} = \min_{\boldsymbol{\alpha}:\boldsymbol{\alpha}\succeq\mathbf{q}_j} \theta_{\boldsymbol{\alpha},j} \geq \theta_{\boldsymbol{\alpha}',j} \geq \theta_{\mathbf{0},j}, \forall\, \boldsymbol{\alpha}' \not\succeq \mathbf{q_j}$.
- For all k and j: $\mathbf{q}_j = \mathbf{e}_k, \theta_{\mathbf{1},j} > \max_{\boldsymbol{\alpha}:\boldsymbol{\alpha}\not\succeq\mathbf{e}_k} \theta_{\boldsymbol{\alpha},j}$.
- After row permutations, the Q-matrix should be of the form

$$\begin{pmatrix} \mathbf{I}_K \\ \mathbf{I}_K \\ \mathbf{Q}' \end{pmatrix},$$

  where $\mathbf{I}_K$ are $K \times K$ identity matrices.
- $\mathbf{Q}'$ satisfies that for $\forall k \in [K]$, there exists item j in $\mathbf{Q}'$, such that $\theta_{\mathbf{e}_k,j} > \theta_{\mathbf{0},j}$.

Intuitively, the first restriction implies that: 1) $\theta_{\boldsymbol{q},j}$ is the probability of responding "1" to item j for all attribute patterns with the required attributes; 2) $\theta_{\boldsymbol{q},j}$ should be greater than any other attribute pattern $\boldsymbol{\alpha}'$ lacking any of the required attributes; and 3) the probability of responding "1" for $\boldsymbol{\alpha} = \mathbf{0}$ should not be greater than that for any other attribute patterns. The second constraint implies that for items requiring only attribute k that the correct response probability for any class missing attribute k is strictly lower than that for any other pattern with attribute k. The third constraint implies that for each attribute, there exists at least two items measuring that attribute only. The fourth constraint indicates there needs to be at least one additional item for each attribute, such that the correct response probability on that item is strictly higher for someone with that attribute only than those with none.

1.2.3 Special Cases

If we regard the RLCMs satisfying the identifiability restrictions above as a general model, many existing CDMs could be seen as cases, thus satisfying the identifiability restrictions. We provide a few examples in this subsection.

A commonly used conjunctive CDM is the DINA model (Junker & Sijtsma, 2001), where examinee i needs to master all attributes required by item j to answer correctly with probability $(1 - s_j)$, and missing any required attributes for this item will result in a correct response probability of $g_j < 1 - s_j$, in other words,

$$P(X_j = 1 \mid \boldsymbol{\alpha}_i) = (1 - s_j)^{\eta_{ij}} g_j^{1-\eta_{ij}}, \tag{1.2}$$

where $\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}}$. Under the DINA model, correct response probability of the all-zero class, along with that of any attribute pattern missing 1 or more required attributes, is g_j, and this probability is strictly lower than $1 - s_j$, the probability of correct response for $\boldsymbol{\alpha}_c \succeq \mathbf{q}_j$. Thus the identifiability constraints of the RLCM are satisfied.

A disjunctive alternative to the DINA model is the deterministic input, noisy-"or"-gate (DINO) model (Templin & Henson, 2006), where only one of the required attributes for item j is needed for the correct response probability of $(1 - s_j)$, and for those without any of the required attributes, the probability of correct response is g_j. In this case $\eta_{ij} = 1 - \prod_{k=1}^{K} (1 - \alpha_{ik})^{q_{ik}}$. We could also verify that the identifiability restrictions are satisfied, with the all-zero class and any pattern lacking any of the required attributes having lowest correct response probability of g_j, and the rest having highest correct response probability for $1 - s_j$.

Other more general models, such as the rRUM (Hartz, 2002), also satisfy the RLCM identifiability constraints. Under the rRUM model, the correct response probability is

$$P(X_j = 1 \mid \boldsymbol{\alpha}_i) = \pi_j^* \prod_{k=1}^{K} r_{jk}^{q_{jk}(1-\alpha_{ik})}.$$

In other words, the probability of responding "1" is π_j^* for someone with all attributes required by the item and a penalty $r_{jk} \in [0, 1]$ is assigned for lacking attribute k if k is required by item j. The all-zero class hence lacks the largest number of required attributes and has lowest correct response probability, the chance of a correct response increases with

the number of mastered attributes required by the item, and it is highest when all required attributes are acquired, satisfying the RLCM identifiability constraints.

More general cognitive diagnosis modeling frameworks were also proposed, where specific models such as DINA, DINO, and the rRUM are special cases. For instance, under the generalized-DINA model (de la Torre, 2011),

$$P(X_j = 1 \mid \boldsymbol{\alpha}_i) = \delta_{j0} + \sum_{\forall k: q_{jk}=1} \delta_{jk}\alpha_{ik} + \sum_{\forall k: q_{jk}=1} \sum_{\forall k': q_{jk'}=1, k'>k} \delta_{jkk'}\alpha_{ik}\alpha_{ik'} + \ldots + \delta_{jkk'k''\ldots} \prod_{k: q_{jk}=1} \alpha_{ik},$$

with $\delta_{j0} > 0$ as the intercept (probability of responding "1" with all-zero attribute pattern), δ_{jk} as the main effects (increase in correct response probability) of having attribute k required by the item, and $\delta_{jkk'\ldots}$ as the interactions between attributes $k, k', \ldots$ required by item j, i.e. the increase in correct response probability by simultaneously mastering these attributes. When the main effects and interactions are free to vary, the G-DINA model is not a special case of the RLCM. However, under the assumption that all the δs are non-negative, the parameters of the G-DINA model are identifiable according to the theorem by Xu (2017).

Another example is the loglinear-CDM (L-CDM, Henson et al., 2009), with

$$\text{logit}P(X_j = 1 \mid \boldsymbol{\alpha}_i) = -\eta_j + \sum_{k=1}^{K} \lambda_{jk}(\alpha_{ik}q_{jk}) + \sum_{k=1}^{K} \sum_{k'>k} \lambda_{jkk'}(\alpha_{ik}q_{jk}\alpha_{ik'}q_{jk'}) + \ldots.$$

One could see that similar to the G-DINA model, the L-CDM also involves the interactions between the attributes required by j in terms of their effects on the probability of responding "1". For identifiability purposes, Henson et al. (2009) suggested imposing a monotonicity assumption, such that mastering any additional skills results in non-decreasing correct response probability. Under this constraint, the L-CDM is a special case of the RLCM, hence the parameters' identifiability is ensured.

1.2.4 Bayesian Formulation

In this subsection we present a Bayesian formulation for RLCMs satisfying Xu's (2017) identifiability restrictions. Under the proposed formulation, the full conditional distributions of the model parameters are analytically tractable, thus enabling the implementation of a Gibbs sampling algorithm for estimation. Previously, Culpepper (2015) introduced a Bayesian formulation of the DINA model allowing the use of Gibbs sampling for parameter estimation. Compared to previous estimation methods with Metropolis-Hastings sampling, where the tuning parameters need to be manually set and adjusted for each data set, Gibbs sampling can demonstrate advantages in both computational efficiency and convenience in applications.

Our formulation for the RLCM assumes that item response function (IRF) follows Equation 1.1. Similar to Culpepper (2015), we assume that the prior probability that examinee i has attribute pattern $\boldsymbol{\alpha}_i$ is

$$P(\boldsymbol{\alpha}_i|\boldsymbol{\pi}) = \prod_{c=1}^{2^K} \pi_c^{\mathcal{I}(\boldsymbol{\alpha}_i=\boldsymbol{\alpha}_c)}, \tag{1.3}$$

where $\mathcal{I}(\cdot)$ is the indicator function, $\pi_c = P(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_c)$ is the probability of having attribute pattern $\boldsymbol{\alpha}_c$ in the population, and the prior for $\boldsymbol{\pi} = \begin{bmatrix} \pi_1 & \dots & \pi_C \end{bmatrix}$ is

$$\boldsymbol{\pi} \sim \text{Dirichlet}(\delta_{01}, \dots, \delta_{0C}), \tag{1.4}$$

where $C = 2^K$.

For the item parameters, we assume that each θ follows a truncated Beta prior distribution. In particular,

$$\theta_{\boldsymbol{\alpha},j} \sim \begin{cases} \text{Beta}(a,b)\mathcal{I}(\theta_{\mathbf{0},j} < \theta_{\boldsymbol{\alpha},j} < \theta_{\boldsymbol{q},j}) & \boldsymbol{\alpha} \not\succeq \mathbf{q}_j, \\ \text{Beta}(a,b)\mathcal{I}(\theta_{\boldsymbol{\alpha},j} = \theta_{\boldsymbol{q},j} > \max_{\boldsymbol{\alpha}' \not\succeq \boldsymbol{q}_j} \theta_{\boldsymbol{\alpha}',j}) & \boldsymbol{\alpha} \succeq \mathbf{q}_j. \end{cases} \tag{1.5}$$

where $\theta_{\boldsymbol{q},j} = \min_{\boldsymbol{\alpha} \succeq \mathbf{q}_j} \theta_{\boldsymbol{\alpha},j}$. Note that the support for the prior distributions of the θ's imposes

Xu's model identifiability constraints.

Full Conditional Distributions

Let $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{iJ})'$ be the observed responses for individual i and $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_N)'$. Then, given the examinees' responses to the J items, $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N)'$, the full conditional distributions of the parameters follow:

1. For $\boldsymbol{\alpha}$:

$$P(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_c \mid \boldsymbol{X}_i, \boldsymbol{\pi}, \boldsymbol{\Theta}) = \frac{\prod_{j=1}^{J} \theta_{\boldsymbol{\alpha}_c,j}^{x_{ij}} (1 - \theta_{\boldsymbol{\alpha}_c,j})^{1-x_{ij}} \pi_c}{\sum_{c=1}^{C} \prod_{j=1}^{J} \theta_{\boldsymbol{\alpha}_c,j}^{x_{ij}} (1 - \theta_{\boldsymbol{\alpha}_c,j})^{1-x_{ij}} \pi_c}. \tag{1.6}$$

2. For $\boldsymbol{\pi}$:

$$\begin{aligned} \boldsymbol{\pi} \mid \boldsymbol{\alpha}, \boldsymbol{\Theta}, \boldsymbol{X} &= \boldsymbol{\pi} \mid \boldsymbol{\alpha} \\ &\sim \text{Dirichlet}(\delta_{01} + \sum_{i=1}^{N} \mathcal{I}(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_1), \ldots, \delta_{0C} + \sum_{i=1}^{N} \mathcal{I}(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_C)). \end{aligned} \tag{1.7}$$

3. For elements of $\boldsymbol{\Theta}$:

$$\theta_{\boldsymbol{\alpha},j} \mid \boldsymbol{\alpha}, \boldsymbol{X}, \boldsymbol{\pi} \sim \tag{1.8}$$

$$\begin{cases} \text{Beta}(a + \sum\limits_{i:\boldsymbol{\alpha}_i = \boldsymbol{\alpha}} X_{ij}, b + \sum\limits_{i:\boldsymbol{\alpha}_i = \boldsymbol{\alpha}} (1 - X_{ij})) \mathcal{I}(\theta_{\mathbf{0},j} < \theta_{\boldsymbol{\alpha},j} < \theta_{\boldsymbol{q},j}), & \text{if } \boldsymbol{\alpha} \not\succeq \mathbf{q}_j; \\ \text{Beta}(a + \sum\limits_{i:\boldsymbol{\alpha}_i \succeq \mathbf{q}_j} X_{ij}, \sum\limits_{i:\boldsymbol{\alpha}_i \succeq \mathbf{q}_j} (1 - X_{ij})) \mathcal{I}(\max_{\boldsymbol{\alpha}' \not\succeq q_j} \theta_{\boldsymbol{\alpha}',j} < \theta_{\boldsymbol{\alpha},j} = \theta_{\boldsymbol{q},j}), & \text{if } \boldsymbol{\alpha} \succeq \mathbf{q}_j. \end{cases}$$

With analytically tractable full conditional distributions of all parameters, a Bayesian Gibbs sampling algorithm could be used for parameter estimation. The Gibbs sampler starts with specifying the initial values, $\boldsymbol{\alpha}_i^{[0]}$ for each i, $\boldsymbol{\pi}^{[0]}$, and $\boldsymbol{\Theta}^{[0]}$. Based on the response matrix, $\mathbf{X}$, the full conditional distributions of $\boldsymbol{\pi}$, $\boldsymbol{\alpha}$'s, and $\boldsymbol{\Theta}$ are then updated in each iteration, and one sample of each is obtained. Specifically, the following procedures are taken in the tth iteration:

- For each subject, obtain $\boldsymbol{\alpha}_i^{[t]}$ from the multinomial distribution, with class probabilities according to equation (4) conditioned upon $\boldsymbol{\Theta}^{[t-1]}$, $\boldsymbol{\pi}^{[t-1]}$, and $\boldsymbol{X}_i$.
- Sample $\boldsymbol{\pi}^{[t]}$ from the Dirichlet full conditional distribution in equation (5) given $\boldsymbol{\alpha}_i^{[t]}$'s.
- For each item j and each attribute class $\boldsymbol{\alpha}_c$ with unique correct response probability, obtain $\theta_{\boldsymbol{\alpha}_c,j}$ from the full conditional distribution in equation (6) based on $\boldsymbol{\alpha}_i^{[t]}$'s and $\boldsymbol{X}$.

The random initial values for θ's are set to satisfy the model constraints for identifiability, and to satisfy the restriction that

$$\max_{\boldsymbol{\alpha}:\boldsymbol{\alpha}\succeq\mathbf{q}_j} \theta_{\boldsymbol{\alpha},j} = \min_{\boldsymbol{\alpha}:\boldsymbol{\alpha}\succeq\mathbf{q}_j} \theta_{\boldsymbol{\alpha},j} = \theta_{\boldsymbol{q},j},$$

within each iteration and for each item. Note that $\theta_{\boldsymbol{q},j}$ was sampled only once and set to the same value for any $\boldsymbol{\alpha} \succeq \mathbf{q}_j$.

1.3 Simulation Studies

The performance of the proposed Gibbs sampler is evaluated in a simulation study in terms of accuracy, efficiency, computational intensity, and convergence. Three factors were considered, generating a total of 12 combinations of factors:

- Number of attributes: $K = 3$ or 4;
- Tetrachoric correlation between attributes: $\rho = 0$ or .5;
- Sample size: $N = 500$, 1000, or 2500.

1.3.1 True parameter generation

We generated 50 replications for each condition using separately generated true $\boldsymbol{\alpha}$'s, $\boldsymbol{\Theta}$, Q-matrix, and response matrix. In each repetition, a random Q-matrix including $J =$

$2 \times (2^K - 1)$ items was generated. To ensure that the randomly generated Q-matrices satisfy the requirements for identifiability, three identity matrices were included in the Q-matrix, so that for each attribute, there are at least 3 items measuring that attribute only. The rest of the rows of the Q-matrix are randomly sampled from possible attribute patterns, except $[0, 0, \ldots, 0]$.

In order to examine the performance of the Gibbs sampler the true item parameters were randomly generated to satisfy the identifiability constraints for the RLCM. More specifically, the following steps were taken to generate the true θ's for each item j:

- Randomly sample $\theta_{\boldsymbol{q},j}$ from Beta$(15, 3)$ and set $\theta_{\boldsymbol{\alpha},j} = \theta_{\boldsymbol{q},j}$ for all $\boldsymbol{\alpha} \succeq \mathbf{q}_j$;
- Randomly sample $\theta_{\mathbf{0},j}$ from truncated Beta$(1.5, 15)$, with the restriction that $\theta_{\mathbf{0},j} < \theta_{\boldsymbol{q},j}$ for $\boldsymbol{\alpha} \succeq \mathbf{q}_j$;
- For each non-zero $\boldsymbol{\alpha}' \not\succeq \mathbf{q}_j$, randomly sample $\theta_{\boldsymbol{\alpha}',j}$ from truncated Beta$(5, 5)$ restricted to the range $(\theta_{\mathbf{0},j}, \theta_{\boldsymbol{q},j})$.

The true attribute patterns of the examinees were randomly generated using similar procedures as Chiu and Köhn (2015). Specifically, for subject i, a K-dimensional multivariate normal latent trait $\boldsymbol{Z}_i$ was first generated from $\mathcal{N}_K(\mathbf{0}, \boldsymbol{\Sigma})$, where the diagonal entries of $\boldsymbol{\Sigma}$ are 1, and the off-diagonal entries are $\rho = 0$ or .5. Next, $\alpha_{ik} = \mathcal{I}(Z_{ik} \geq \Phi^{-1}(\frac{k}{K+1}))$ was obtained for each k, giving the true attribute pattern of subject i, $\boldsymbol{\alpha}_i$.

1.3.2 Parameter Estimation

The prior distribution of the population membership probabilities, $\boldsymbol{\pi}$, was set to be Dirichlet$(\mathbf{1})$, and each of the correct response probabilities, $\theta_{\boldsymbol{\alpha},j}$, was assigned to follow prior distribution of Beta$(1, 1)$. To start the Gibbs sampler, initial values were assigned to $\boldsymbol{\Theta}$, $\boldsymbol{\alpha}_i$'s, and $\boldsymbol{\pi}$. In particular, the initial item parameters, $\theta^{[0]}_{\boldsymbol{\alpha},j}$, were randomly sampled following the same

procedures of generating the true item parameters, so that they satisfy the identifiability constraints for the RLCM. Each entry of the initial attribute patterns for the examinees, $\boldsymbol{\alpha}_i^{[0]}$, for each individual i, was randomly sampled from Bernoulli(.5). The initial values of the population membership probabilities, $\boldsymbol{\pi}^{[0]}$, were generated from the Dirichlet distribution with $\delta_{01} = \ldots = \delta_{0C} = 1$. Then, the MCMC algorithm was executed following the procedures formerly described.

The algorithm terminates after all T iterations, and the last $10,000$ iterations were chosen as the post burn-in samples for parameter estimation. The expected a posteriori (EAP) estimate of the item parameters and the population proportions, as well as the maximum a posteiori (MAP) estimate of the individuals' attribute patterns, were taken as the point estimates of the item, population membership, and examinee attribute pattern parameters. The Gibbs sampling algorithm mentioned above for parameter estimation were written in C++ and were deployed in R via the Rcpp package (Eddelbuettel et al., 2011).

To determine the number of iterations needed for the Markov chain to converge, T, we first evaluated the convergence as a function of number of iterations for the most computationally intensive condition, in other words, $N = 2500, K = 4$, and $\rho = .5$. Based on the same response matrix, 5 separate chains of the Gibbs Sampler were run, with different starting values. For both $\boldsymbol{\Theta}$ and $\boldsymbol{\pi}$, the Brooks-Gelman multivariate scale reduction factor, $\hat{R}$ (Brooks & Gelman, 1998), was computed for chain lengths of 20000 to 50000 at increments of 500, with 10000 post-burn-in iterations. The code for computing the Brooks-Gelman diagnostic statistic can be found in the `coda` package in R (Plummer, Best, Cowles, & Vines, 2006). Figure 1.1 shows the progression of the $\hat{R}$ statistic as a function of the number of iterations for MCMC sampling. The multivariate $\hat{R}$ statistic decreased to around 1.2 after approximately 35000 iterations. Thus, the simulation study used 35000 MCMC cycles with 25000 as the burn-in.

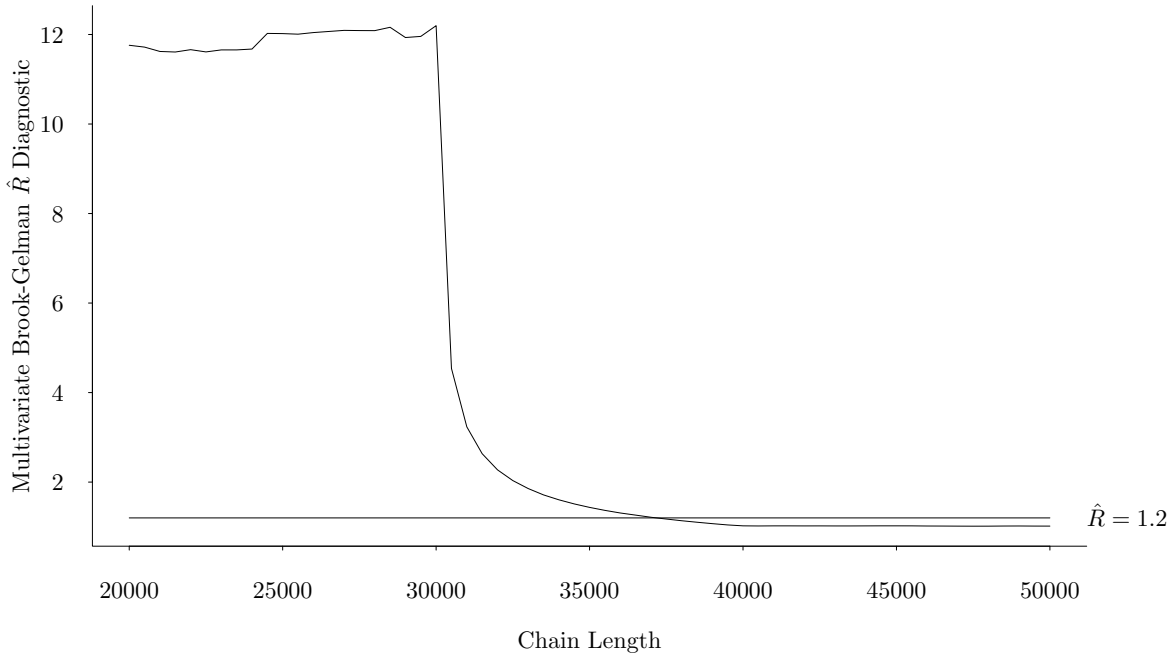

Figure 1.1: Brooks-Gelman $\hat{R}$ for the convergence of $\boldsymbol{\Theta}$ and $\boldsymbol{\pi}$ in 5 chains, under the condition of $N = 2500, K = 4, \rho = .5$. Horizontal line represents the $\hat{R} = 1.2$ cutoff that is commonly used as a threshold for parameter convergence.

1.3.3 Evaluation Criteria

We evaluate the performance of the proposed Gibbs sampling algorithm in terms of parameter recovery and computational efficiency. For each distinct true item parameter in each repetition, $\theta_{\boldsymbol{\alpha},j}$, we calculate the bias and root-mean-square deviation of the parameter estimate using

$$\text{Bias}(\hat{\theta}_{\boldsymbol{\alpha},j}) = \hat{\theta}_{\boldsymbol{\alpha},j} - \theta_{\boldsymbol{\alpha},j},$$

$$\text{RMSE}(\hat{\theta}_{\boldsymbol{\alpha},j}) = \sqrt{\frac{\sum_{t=T_{burn}+1}^{T}(\hat{\theta}_{\boldsymbol{\alpha},j}^{[t]} - \theta_{\boldsymbol{\alpha},j})^2}{T - T_{burn}}},$$

where $\hat{\theta}_{\boldsymbol{\alpha},j}$ is the EAP estimate of the item parameter, and $\hat{\theta}_{\boldsymbol{\alpha},j}^{[t]}$ is the sample for the item parameter in the tth iteration of the MCMC algorithm.

Similarly, for each repetition, the bias and RMSE were computed for each population

probability estimate, $\hat{\boldsymbol{\pi}}$. We thus computed the bias and RMSE of the $\boldsymbol{\pi}$ samples obtained from the MCMC cycles using the values implied by the multivariate probability as the true $\boldsymbol{\pi}$.

The estimation accuracy of the examinees' attribute patterns were evaluated in terms of the Attribute-wise Agreement Rate (AAR) and the Pattern-wise Agreement Rate (PAR), given by

$$AAR = \frac{\sum_{i=1}^{N}\sum_{k=1}^{K}\mathcal{I}(\alpha_{ik} = \hat{\alpha}_{ik})}{N \times K}, \text{and}$$

$$PAR = \frac{\sum_{i=1}^{N}\mathcal{I}(\boldsymbol{\alpha}_i = \hat{\boldsymbol{\alpha}}_i)}{N}.$$

Finally, to evaluate the computational efficiency of the proposed algorithm, for each simulation condition, the computation time for 35000 iterations of the MCMC algorithm on a PC laptop with Intel Core i7 CPU 2.60 GHz processor and 8GB RAM was recorded.

1.3.4 Results

In Figures 1.2 and 1.3, we present the bias and RMSE of item parameter estimates when $K = 3$. Results for different sample sizes and attribute correlations are shown in separate grids. The x-axis on each plot represents the Q-matrix loadings, $\mathbf{q}_j$, of an item, where the dots, from bottom to top, denote attribute $1, 2$, and 3, respectively, with a solid black dot indicating the item requiring that attribute (hollow otherwise). The y-axis in Figure 1.2 represents the averaged bias of item parameter estimates, where across repetitions and across all items with same Q-matrix loadings, the bias for elements of $\widehat{\boldsymbol{\Theta}}$ are collapsed for each $\boldsymbol{\alpha}_c$ configuration. Similarly for the RMSEs in Figure 1.3. As we have mentioned above, for each item, the correct response probabilities, elements of $\widehat{\boldsymbol{\Theta}}$ can be categorized into three groups based on the type of skill mastery pattern, namely people who have not acquired any attributes (i.e., $\boldsymbol{\alpha} = \mathbf{0}$), people who have not acquired all skills required by item j (i.e., $\boldsymbol{\alpha} \succ \mathbf{0}$ and $\boldsymbol{\alpha} \not\succeq \mathbf{q}_j$), and people who have acquired all skills required by the item (i.e., $\boldsymbol{\alpha} \succeq \mathbf{q}_j$). These three types of item parameters for each item are represented with different

shapes on the bias and RMSE plots (cross, hollow dot, and solid dot, respectively).

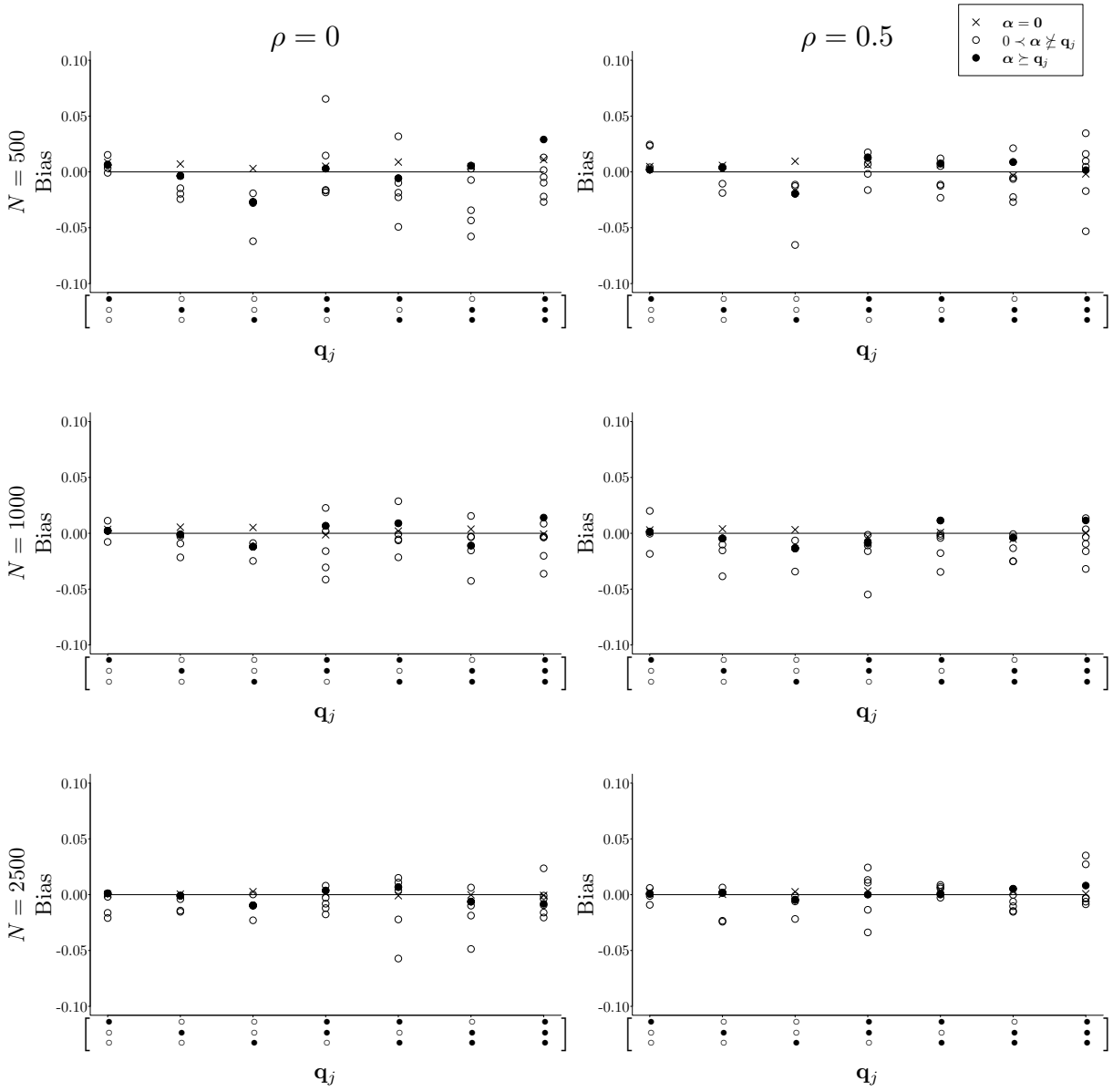

Figure 1.2: Bias of item parameter (i.e., elements of $\boldsymbol{\theta}$) estimates across repetitions when the number of attributes is $K = 3$. *Note.* The three different plotting characters item parameters for three groups based on the type of skill mastery pattern: a cross denotes parameters for the class without any attributes (i.e., $\boldsymbol{\alpha} = \mathbf{0}$); a hollow dot represents classes that do not have all required attributes (i.e., $\boldsymbol{\alpha} \succ \mathbf{0}$ and $\boldsymbol{\alpha} \not\succeq \mathbf{q}_j$); and a solid dot indicates classes that have all requisite skills (i.e., $\boldsymbol{\alpha} \succeq \mathbf{q}_j$).

Overall, the proposed estimation procedures have accurately recovered the true item parameters across all of the simulation conditions, with bias close to zero and RMSE below

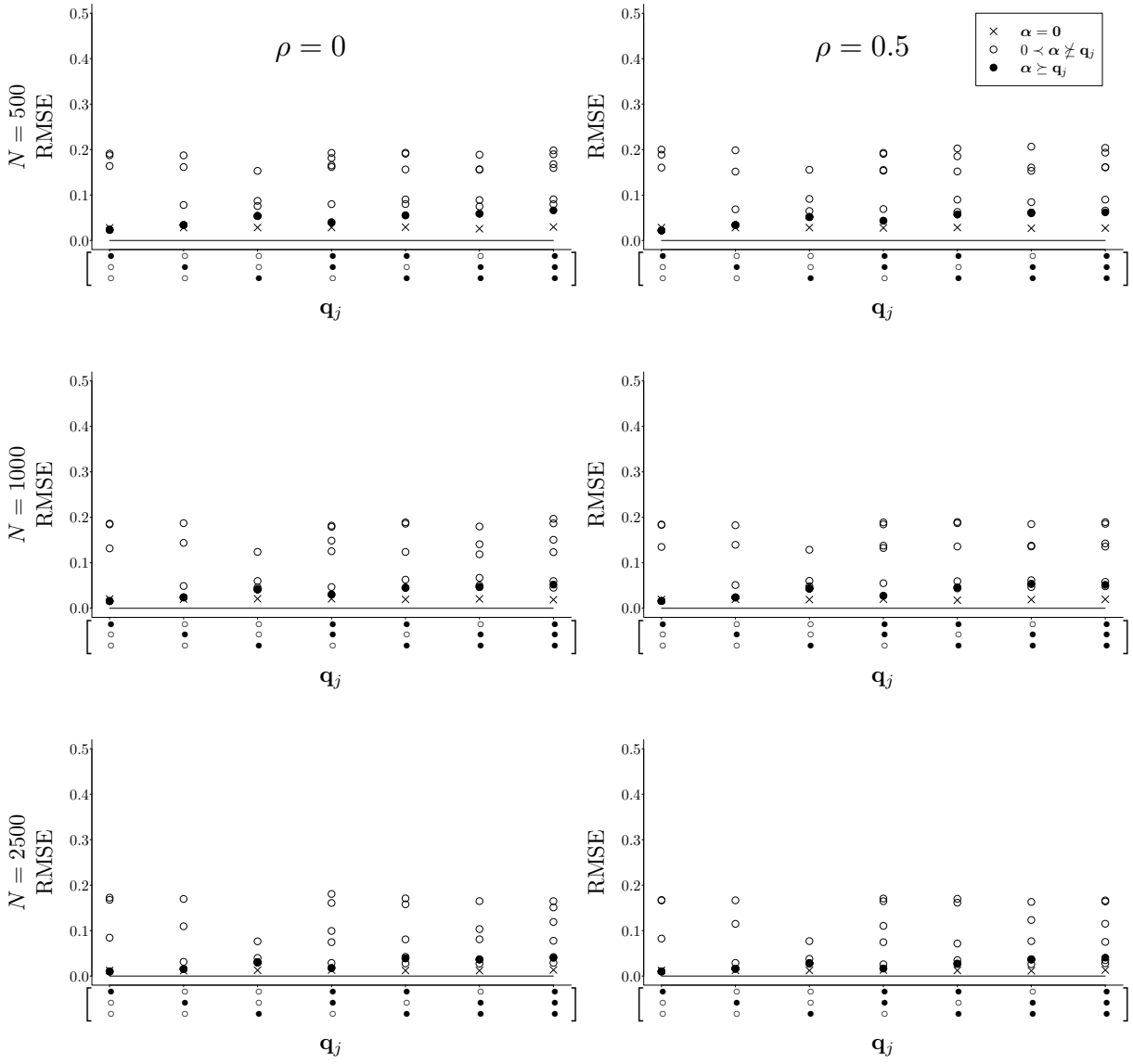

Figure 1.3: RMSE of item parameter (i.e., elements of $\boldsymbol{\theta}$) estimates for $K = 3$. *Note.* The three different plotting characters item parameters for three groups based on the type of skill mastery pattern: a cross denotes parameters for the class without any attributes (i.e., $\boldsymbol{\alpha} = \mathbf{0}$); a hollow dot represents classes that do not have all required attributes (i.e., $\boldsymbol{\alpha} \succ \mathbf{0}$ and $\boldsymbol{\alpha} \not\succeq \mathbf{q}_j$); and a solid dot indicates classes that have all requisite skills (i.e., $\boldsymbol{\alpha} \succeq \mathbf{q}_j$).

0.2 for almost all elements of $\widehat{\boldsymbol{\Theta}}$. Whereas the averaged bias and RMSE for item parameters associated with $\boldsymbol{\alpha}_c \succeq \mathbf{q}_j$ and $\boldsymbol{\alpha}_c = \mathbf{0}$ were all very close to 0, a few item parameter estimates associated with $\mathbf{0} \prec \boldsymbol{\alpha}_c \not\succeq \mathbf{q}_j$ had larger deviations from the true values and larger RMSEs.

After manually inspecting the item parameter bias of a few repetitions, we infer that this is likely because some of the $\boldsymbol{\alpha}$ classes, such as $\boldsymbol{\alpha} = [0, 0, 1]$ and $[0, 1, 1]$, have very small sample sizes due to how we generated the true $\boldsymbol{\alpha}$. Consistent with previous studies we observe smaller bias and RMSE for items with simple Q-matrix loadings. A slight decrease in bias and RMSE was seen as we increased the sample size. However, the estimation accuracy and efficiency of certain item parameters with $\boldsymbol{\alpha}_c \not\succeq \mathbf{q}_j$ did not improve significantly with larger sample sizes, as the attribute classes with near zero implied membership probabilities still have expected sample size below 25 even for $N = 2500$. The bias in item parameter estimates seemed slightly lower when $\rho = .5$ than when the attributes are independent.

Figures 1.4 and 1.5 present the bias and RMSE of $\hat{\boldsymbol{\Theta}}$ when $K = 4$. We note here that the number of items J, was equal to $2 \times (2^K - 1)$ and hence varied across different K's, thus the results from the $K = 4$ condition is not directly comparable to when $K = 3$. When $K = 4$, the estimation algorithm demonstrated similar performance to when $K = 3$. One difference we could observe from the figures is that when $K = 4$, the item parameters associated with $\boldsymbol{\alpha}_c \succeq \mathbf{q}_j$, represented by solid dots, tend to be consistently overestimated, with positive bias between 0 and .05. This positive bias is more obvious for items with complex Q-matrix loadings and for small sample sizes, and it gradually fades away as the sample size increases.

The bias and RMSE of the estimation of population membership probabilities, $\boldsymbol{\pi}$, are presented in Figures 1.6 and 1.7 for $K = 3$, and in Figures 1.8 and 1.9 for $K = 4$. The x-axis represents the attribute pattern, $\boldsymbol{\alpha}_c$, and from bottom to top, the dots represent attributes 1 to K, with hollow dots indicating $\alpha_k = 0$ (solid otherwise). The y-axis denotes the averaged-across-repetitions bias of $\boldsymbol{\pi}$ estimates associated with that attribute pattern. Under all simulation conditions we observed bias and RMSE of the $\boldsymbol{\pi}$ estimates close to zero, suggesting high recovery of the true population membership probabilities. When $K = 3$ and $N = 500$, the absolute bias and RMSE for $\hat{\pi}_c$ was larger for $\boldsymbol{\alpha}_c = [1, 0, 0]$ and $\boldsymbol{\alpha}_c = [1, 1, 0]$, where the averaged bias equaled $-.05$. This is likely due to the low implied membership

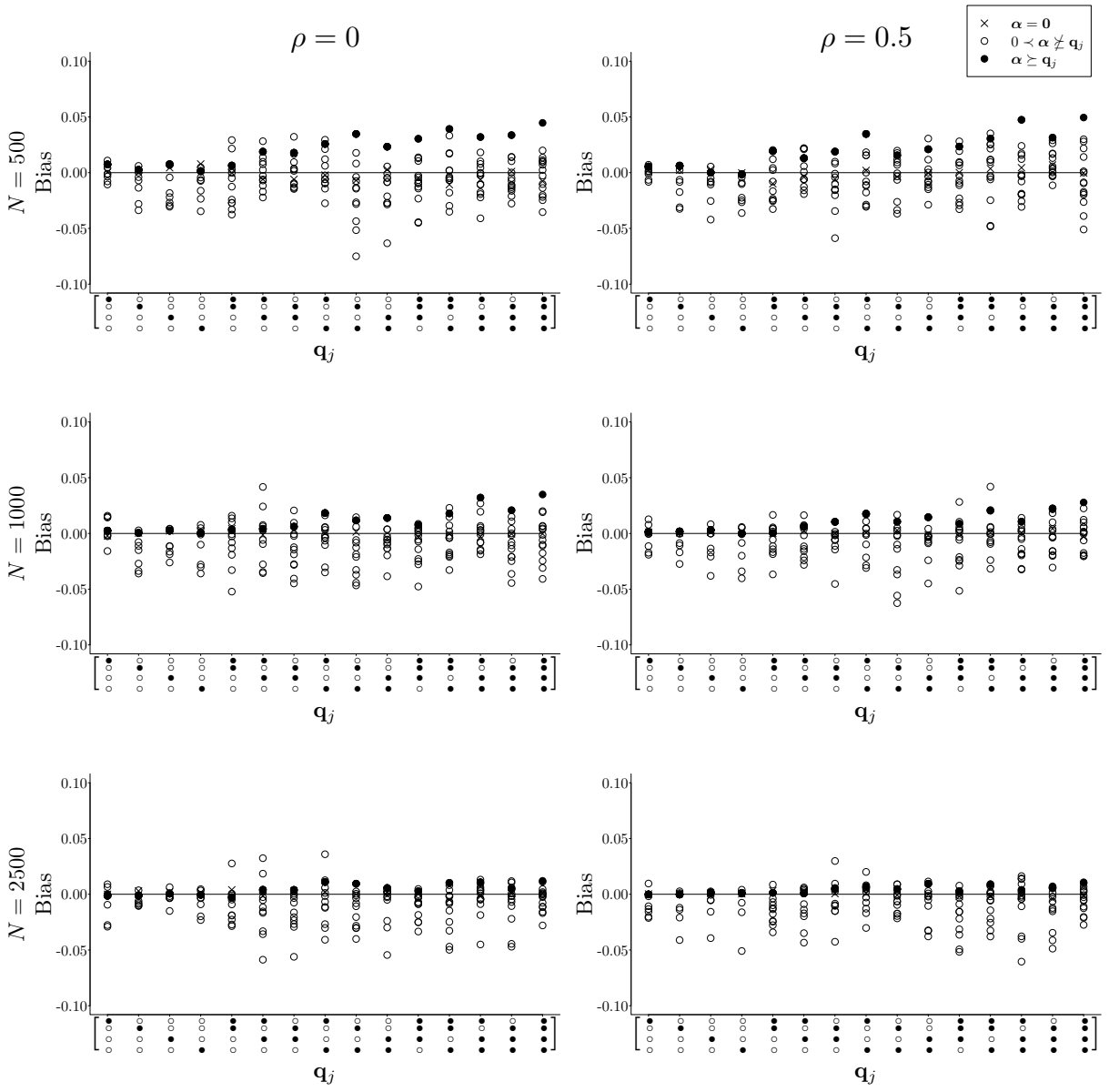

Figure 1.4: Bias of item parameter (i.e., elements of $\boldsymbol{\theta}$) estimates for $K = 4$. *Note.* The three different plotting characters item parameters for three groups based on the type of skill mastery pattern: a cross denotes parameters for the class without any attributes (i.e., $\boldsymbol{\alpha} = \mathbf{0}$); a hollow dot represents classes that do not have all required attributes (i.e., $\boldsymbol{\alpha} \succ \mathbf{0}$ and $\boldsymbol{\alpha} \not\succeq \mathbf{q}_j$); and a solid dot indicates classes that have all requisite skills (i.e., $\boldsymbol{\alpha} \succeq \mathbf{q}_j$).

probabilities for these two patterns, both of which are below .01. However, as sample size increases and in the case of $K = 4$ (with more items), this bias gets close to zero. Similar to item pameter estimation, we observed smaller bias and lower RMSE for $\hat{\pi}_c$ when $\boldsymbol{\alpha}_c$ is "1"

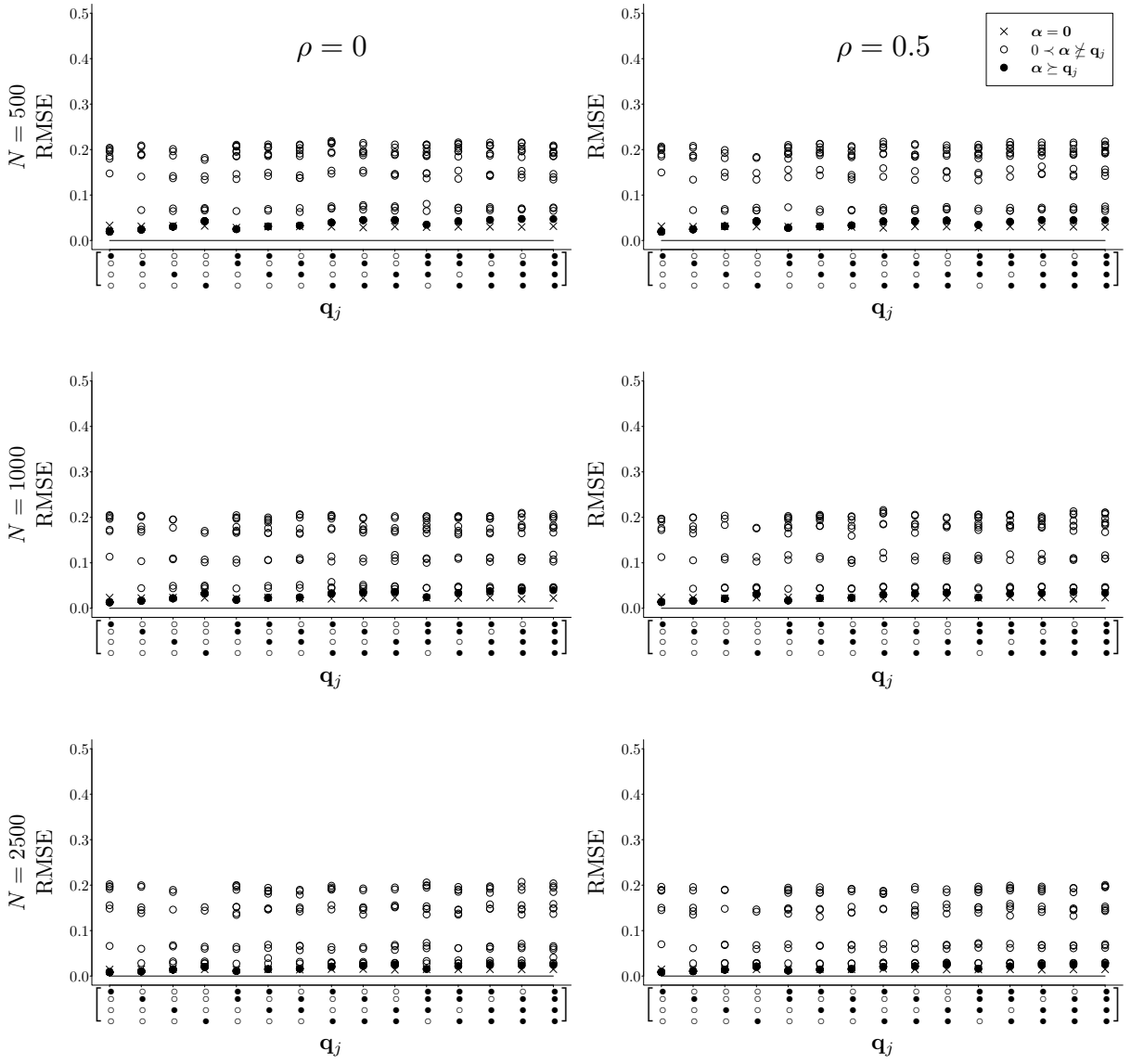

Figure 1.5: RMSE of item parameter (i.e., elements of $\boldsymbol{\theta}$) estimates for $K = 4$. *Note.* The three different plotting characters item parameters for three groups based on the type of skill mastery pattern: a cross denotes parameters for the class without any attributes (i.e., $\boldsymbol{\alpha} = \mathbf{0}$); a hollow dot represents classes that do not have all required attributes (i.e., $\boldsymbol{\alpha} \succ \mathbf{0}$ and $\boldsymbol{\alpha} \not\succeq \mathbf{q}_j$); and a solid dot indicates classes that have all requisite skills (i.e., $\boldsymbol{\alpha} \succeq \mathbf{q}_j$).

on one attribute only. And finally, we did not find any consistent differences for $\rho = 0$ and $\rho = .5$.

The forth and fifth columns of Table 1 present the accuracy of attribute pattern $\boldsymbol{\alpha}_i$ es-

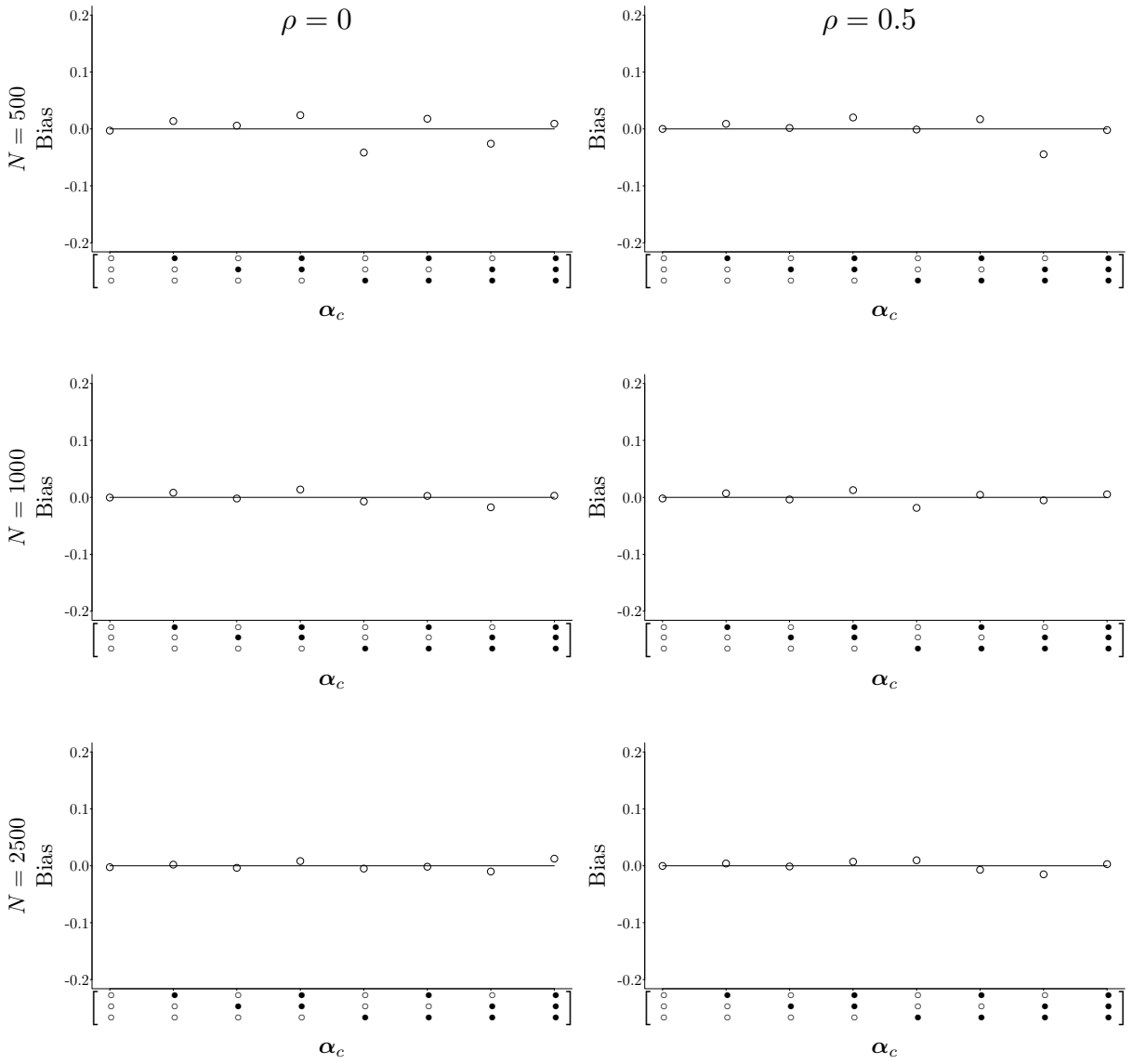

Figure 1.6: Bias for population membership probability ($\boldsymbol{\pi}$) estimates when $K = 3$.

timates for the simulated respondents, in terms of averaged-across-repetitions attribute-wise agreement rate (AAR) and pattern-wise agreement rate (PAR). The proposed estimation routine for the RLCM achieved over 87% of attribute entry recovery and over 70% of whole pattern recovery of the simulated subjects for $K = 3$. And for $K = 4$, because the number of items is larger, the accuracy was even higher, achieving over 90% for attribute entries recovery and over 72% for pattern recovery across all simulation conditions. We also observed

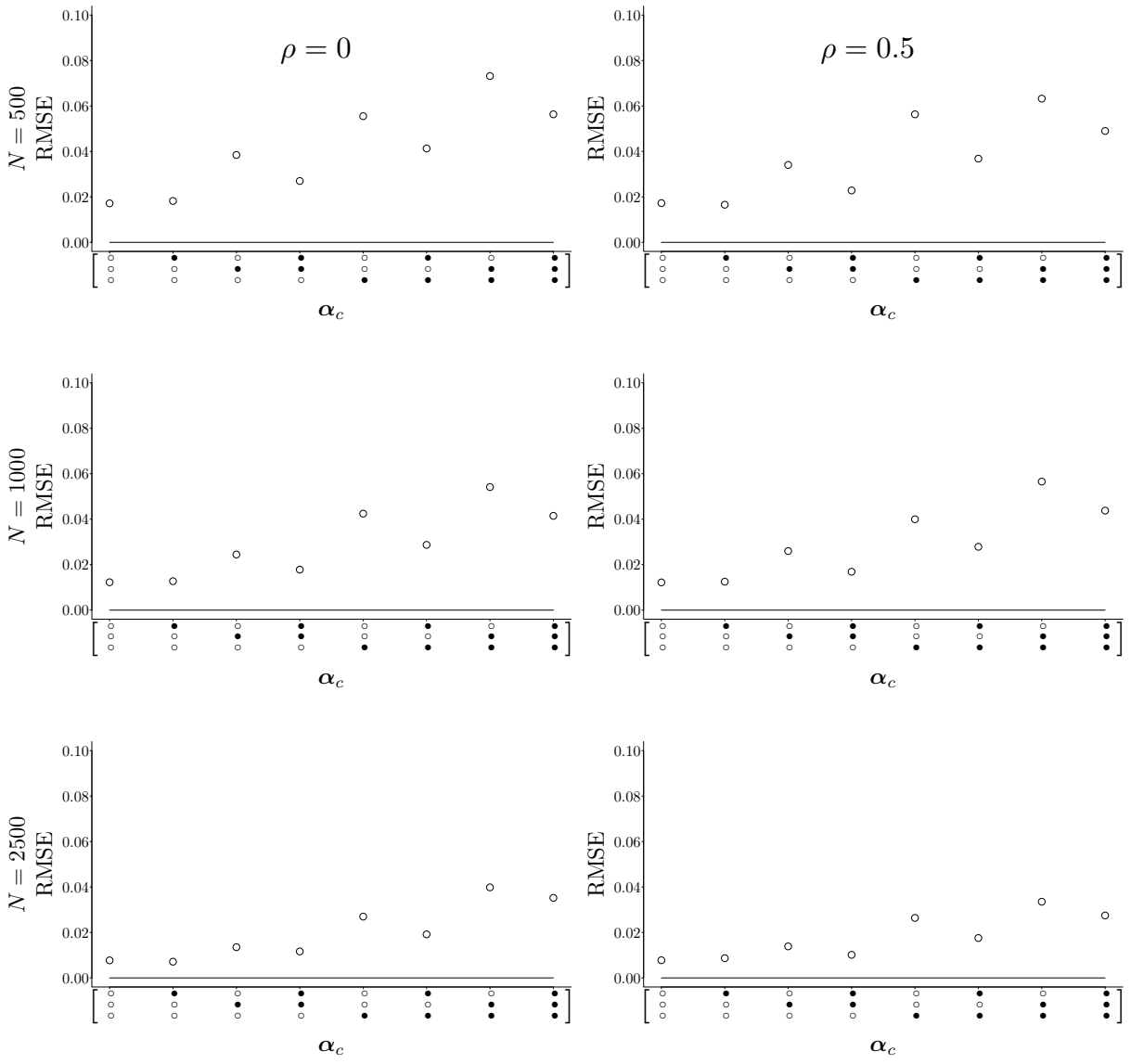

Figure 1.7: RMSE for population membership probability ($\boldsymbol{\pi}$) estimates when $K = 3$.

that the AAR and PAR were higher for larger sample sizes, which is potentially due to the increase in item parameter estimation accuracy, and when the attributes are correlated.

The last column of Table 1.1 summarizes the computation time in seconds for 35000 MCMC chains in each of the simulation conditions, on a personal laptop with Intel i7-4510U CPU and 8GB RAM. We can see that for the simpler conditions, such as when $K = 3$ and $N = 500$ or 1000, the algorithm terminated after slightly more than 2 minutes. The

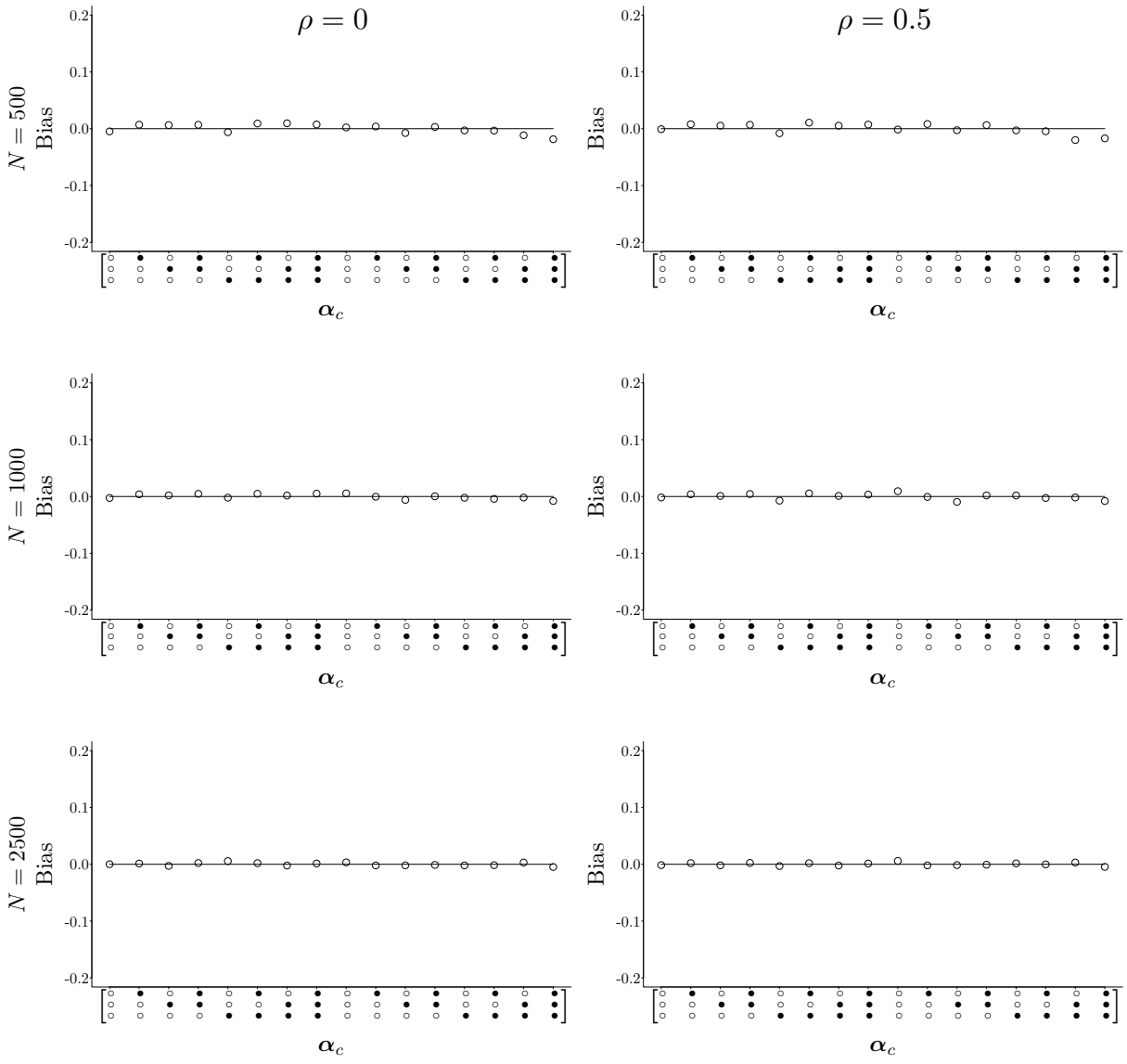

Figure 1.8: Bias for population membership probability ($\boldsymbol{\pi}$) estimates when $K = 4$.

computation time increases as the sample size, item number, and number of attributes increase, but even for the most computational intensive condition, where $K = 4, N = 2500$, and $J = 30$, the algorithm terminated after around 44 minutes. Given the flexibility of the model and the large number of free parameters to be estimated, we believe that the proposed Gibbs sampler could accurately recover the underlying model parameters with a manageable computation time.

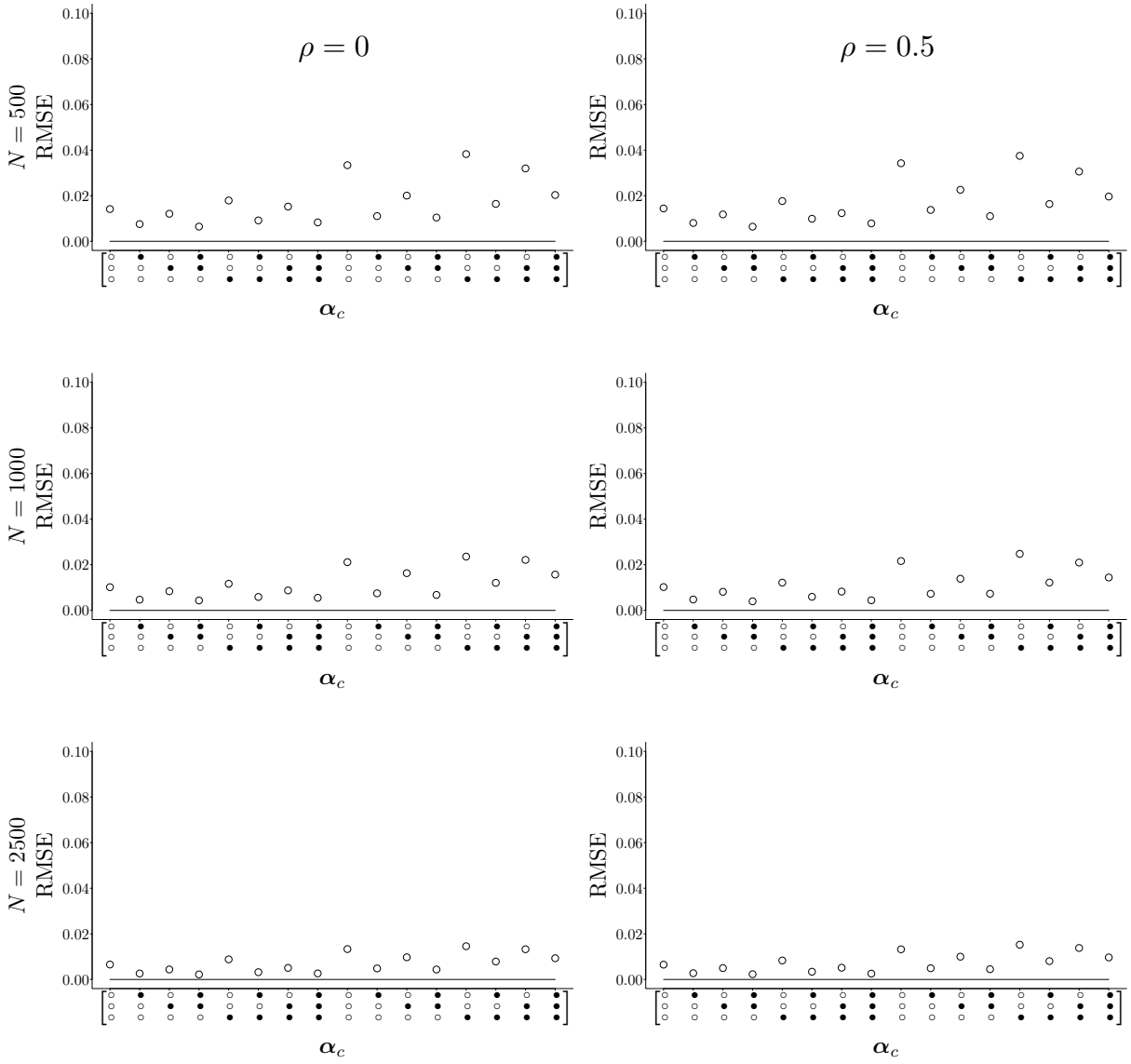

Figure 1.9: RMSE for population membership probability ($\boldsymbol{\pi}$) estimates when $K = 4$.

1.4 Application: Examination for the Certificate of Proficiency in English

In this section we present the results for fitting the RLCM to the Examination for the Certificate of Proficiency in English (ECPE) data set (Henson & Templin, 2007; Templin & Hoffman, 2013), which could be obtained from the `cdm` package in R (George, Robitzsch,

Table 1.1: Attribute-wise Agreement Rate (AAR), Pattern-wise Agreement Rate (PAR), and computational Time of the RLCM Gibbs Sampler for Values of K, ρ, and N.

K	ρ	N	AAR	PAR	Time (sec)
3	0	500	0.876	0.701	125.22
3	0.5	500	0.882	0.714	129.98
3	0	1000	0.888	0.726	243.66
3	0.5	1000	0.889	0.729	132.63
3	0	2500	0.893	0.738	463.35
3	0.5	2500	0.895	0.744	468.66
4	0	500	0.905	0.721	693.81
4	0.5	500	0.905	0.724	704.82
4	0	1000	0.917	0.755	1101.42
4	0.5	1000	0.920	0.766	1187.02
4	0	2500	0.924	0.779	2499.27
4	0.5	2500	0.924	0.780	2641.05

Kiefer, Groß, & Ünlü, 2016). The ECPE data was collected from a large-scale English language assessment using cognitive diagnosis models, and it contains 2922 examinees' dichotomous responses to 28 items. Three attributes, namely morphosyntactic rules, cohesive rules, and lexical rules, and the 28×3 Q-matrix were identified by content experts.

We fit both the RLCM and the rRUM (Hartz, 2002) to the ECPE response data. The rRUM model is a multiplicative model, where lacking each required attribute introduces a "penalty" to the probability of correct response. The rRUM satisfies Xu's (2017) identifiability conditions and can be considered as a more parsimonious model nested in the RLCM. By comparing the performance of rRUM and RLCM in fit to the empirical data, we can assess the extent to which the introduction of additional parameters could improve model fit.

Similar to the simulation studies, 35000 iterations of the MCMC algorithm was used for the RLCM, and 20000 iterations was used for the rRUM, as previous studies (Culpepper & Hudson, 2017) suggested that this would be sufficient for convergence when $K = 3$. For both models, 10000 post-burn iterations were used for estimating the model parameters. We compare the fit of the two models in two main aspects, specifically overall model fit and item fit.

The overall fit of the two models to the ECPE data was evaluated in terms of the posterior predictive probabilities of the observed total scores (i.e., number of correct responses). For both models, using each MCMC iterations' samples of attribute patterns and item parameters, we simulated a response matrix and computed the total score of the respondents in that iteration. The percentile rank of the observed total score relative to the simulated total scores across the 10000 iterations was then obtained for each subject, used as the posterior predictive probability (PPP) of the observed total score. Figure 1.10 presents the PPPs for observed scores under the two models. The x-axis denotes the observed total score of the respondent, and the y-axis gives the PPP of the observed score, relative to the simulated ones over 10000 iterations. Each point in the plot represents one of the 2922 subjects, and the PPPs obtained under the two models are plotted in separate panels, with the RLCM on the left and the rRUM on the right. Because the PPPs indicate the percentile rank of the observed total score to the simulated ones based on the sampled model parameters in the MCMC algorithms, the more extreme the PPPs (i.e., closer to 0 or 1), intuitively the larger the distance between the observed data and the model-predicted data. Overall, the percentage of extreme PPPs above .95 or below .05 was slightly lower for the RLCM (7.67%) than for the rRUM (9.86%). We could observe from the figure that the PPPs for the rRUM and for the RLCM are very similar in the middle range of total scores, but for respondents with total scores less than 14 or more than 24, the PPPs under the rRUM appeared more extreme than under the RLCM, which indicates better fit of the RLCM for participants with more extreme observed scores. It should be noted that neither models fit well for participants with extreme low or extreme high scores, with PPPs of both models around 0 and 1, respectively.

For item fit, we first examined the congruence between observed and expected proportion correct in each equivalence class, as suggested in Sinharay and Almond (2007) for assessing item fit in cognitive diagnosis models. Sinharay and Almond (2007) defined equivalence classes in terms of partitions of attribute types, so that in each equivalence class, all attribute types have the same probability of correct response by the model. For the RLCM,

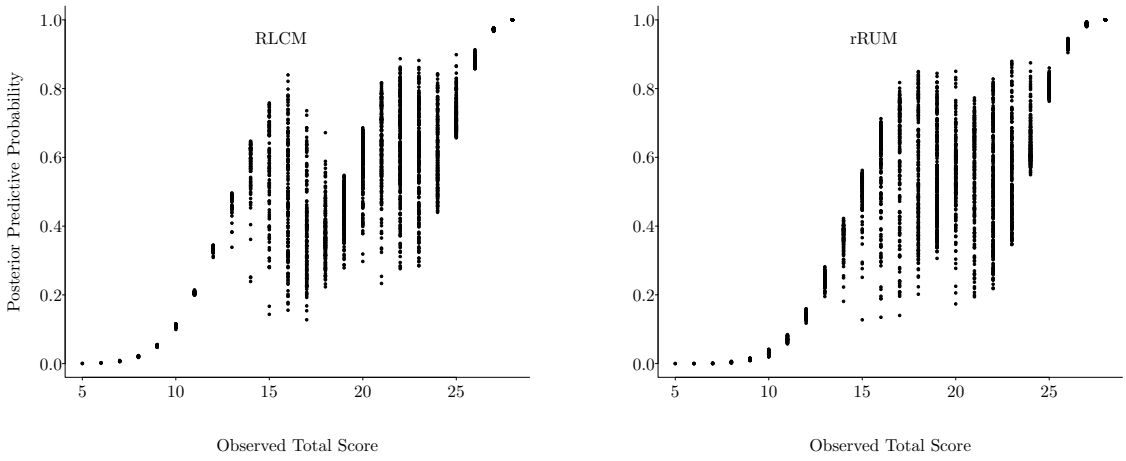

Figure 1.10: Posterior predictive probabilities of observed total scores for the RLCM and the rRUM.

all attribute patterns $\boldsymbol{\alpha}_c \succeq \mathbf{q}_j$ are required to have the same correct response probability on an item, thus these patterns form an equivalence class. None of the other patterns are required by the model to have the same correct response probability, thus each of them form their own class. Under the rRUM model, individuals with the same α_{ik} values on attributes required by the item, i.e. $\{k : q_{jk} = 1\}$, have the same correct response probability, forming equivalence classes. Ideally, item fit can be evaluated by comparing the observed proportion correct and the expected proportion correct estimated from the model for each equivalence class. Because the true attribute patterns of the respondents are unknown in practice, Sinharay and Almond (2007) suggested using the MCMC-sampled attribute patterns in each iteration as an approximate. For both the RLCM and rRUM, we used the sampled attribute patterns of the examinees to calculate the "observed" proportion correct in each equivalence class for each item in that iteration. A 95% credible interval band over iterations was then obtained for each item and each equivalence class' observed proportion correct. This band is then compared to the expected proportion correct given by the EAP of item parameters. An expected proportion correct near the two poles of or outside the

95% band would indicate a discrepancy between the observed data and model predictions for that item and class.

Figures 1.11 and 1.12 present the model predicted proportion correct and the MCMC-sampled observed proportion correct of each item and equivalence class. On each item plot, the x-axis denotes the attribute pattern and the y-axis gives the proportion correct for that pattern. The dots represent the expected proportion correct computed based on the EAP of item parameters, and we can see that attribute patterns that fall into the same equivalence class have the same model predicted proportion correct. The vertical bands represent the 95% band for MCMC-sample observed proportion correct, and because some of the band widths were very close to 0, they are not shown on the figures. Again, patterns from the same equivalence class have the same aggregated proportion correct band, and since the rRUM has a smaller number of equivalence classes, the number of subjects in each class in the MCMC-cycles are larger than that of the RLCM, resulting in narrower bands. Comparing the plots for the RLCM and the rRUM, we observe that the predicted proportions correct differed for some classes associated with $\boldsymbol{\alpha} \not\succeq \mathbf{q}_j$. For example, on item 19, the RLCM predicts only the all-zero class has expected proportion correct around .4, and the other attribute patterns lacking the required skill 3 were expected to have proportion correct close to .6 or .8. The rRUM, on the other hand, predicts that all attribute patterns lacking skill 3 have expected proportion correct near .4. Looking at the positions of the expected proportions correct relative to the 95% observed band, we see that all dots are close to the center of the bands for both models, thus both models did well in predicting the first moments (proportion correct) of the items.

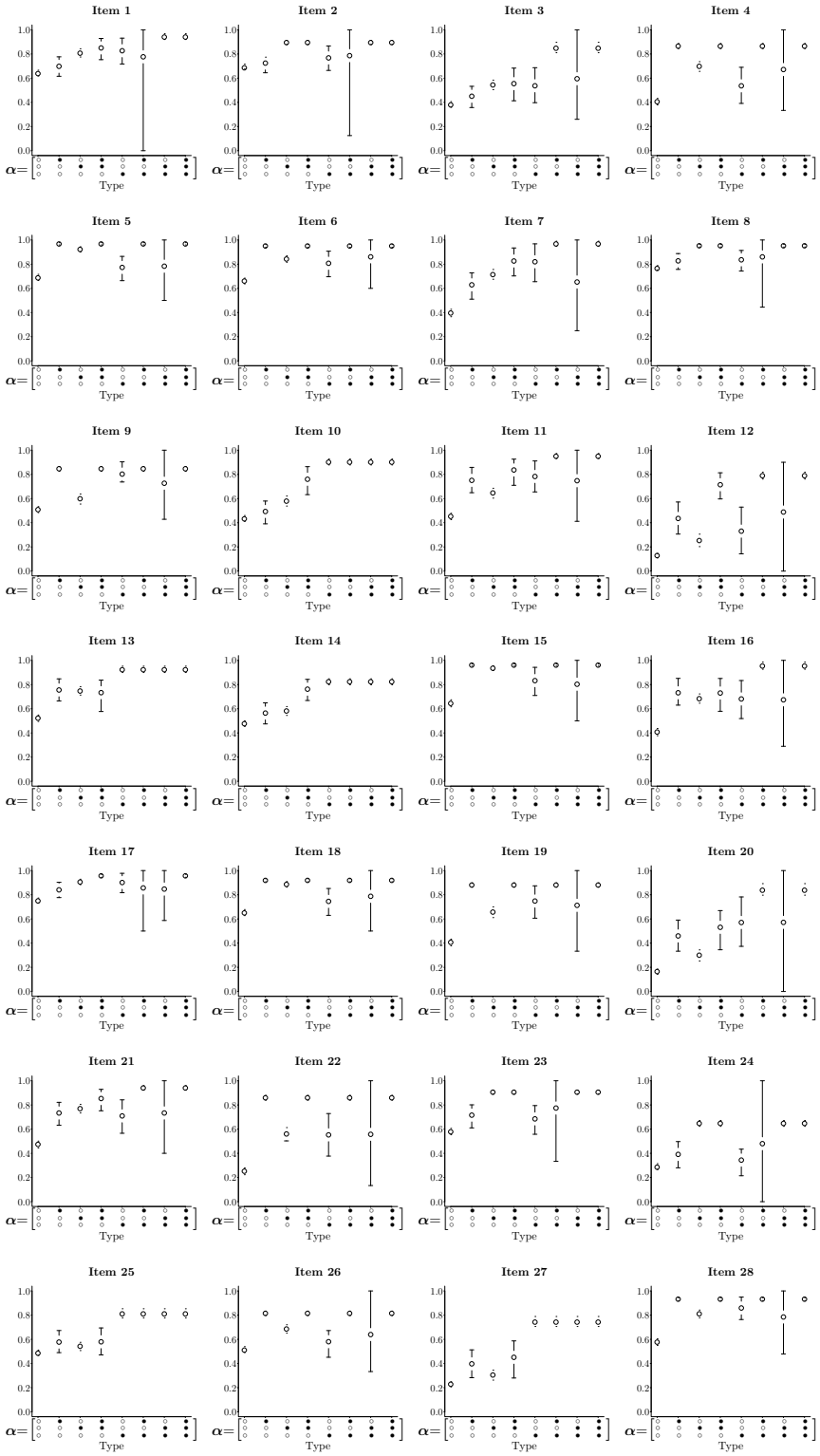

Figure 1.11: Plots of RLCM empirical item response probabilities across attribute profiles $\boldsymbol{\alpha}_c$. Note. Some of the band widths were very close to 0, they are not shown on the figures.

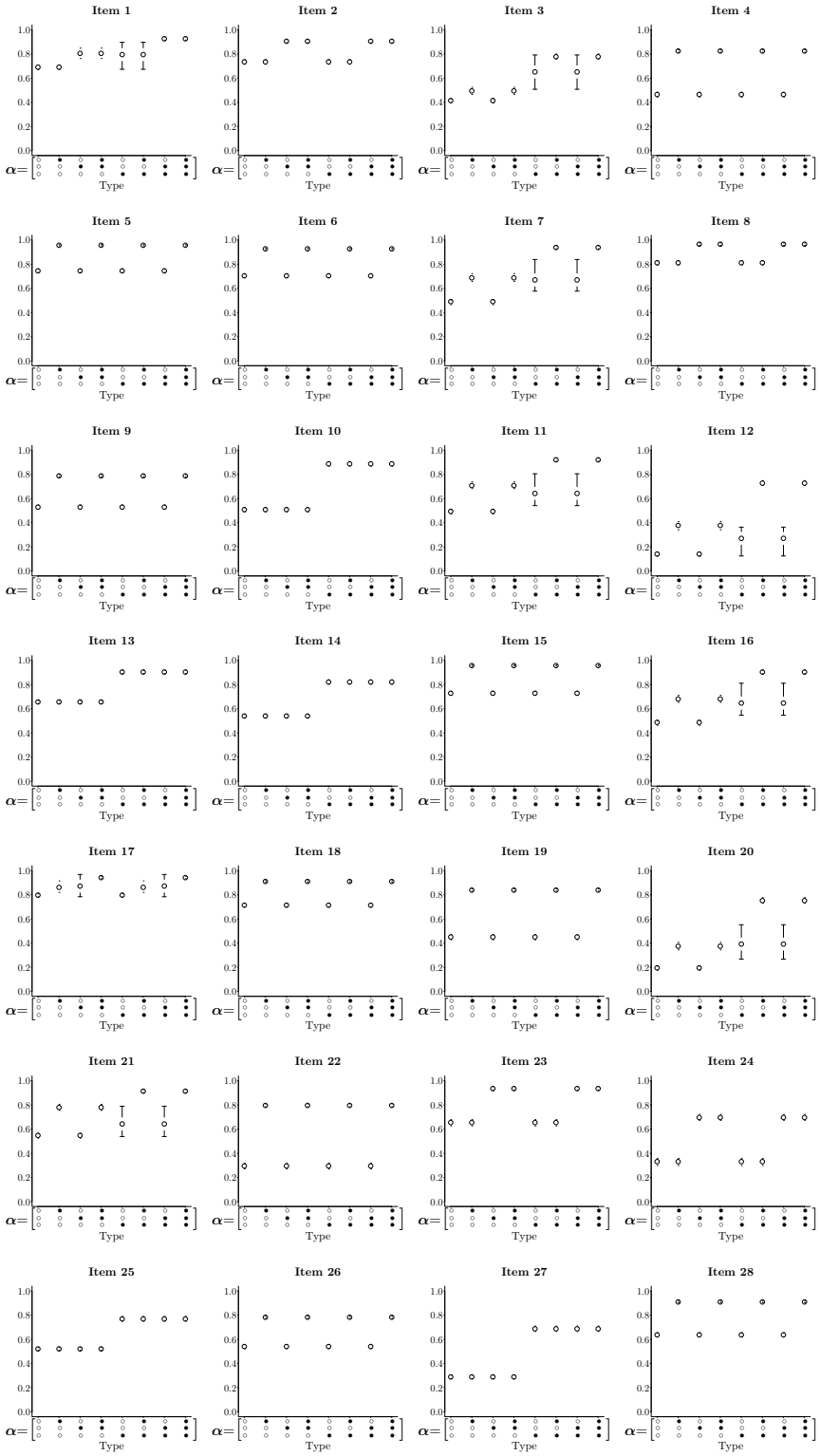

Figure 1.12: Plots of rRUM empirical item response probabilities across attribute profiles $\boldsymbol{\alpha}_c$. Note. Some of the band widths were very close to 0, they are not shown on the figures.

To further evaluate item-level fit, we examined the item-pairwise odds ratios (ORs). For any item pair, Sinharay, Johnson, and Stern (2006) suggested using $OR = (n_{11}n_{00})/(n_{01}n_{10})$, where n_{11} is number of respondents responding to both items correctly, n_{01} is number of respondents who answered item 1 wrong and item 2 correctly, etc., as a measure of item-pairwise associations. The OR would then be used for posterior predictive model checking (PPMC). Similar to the PPMC for total scores, for each MCMC iteration we simulated a 2922×28 response matrix based on the sampled parameters. Each response matrix was used to obtain the ORs for each item pair in that iteration. The observed item-pairsise ORs were calculated using the observed ECPE response matrix, and the posterior predictive probability of the observed OR for each item pair among the 10000 model predicted ORs was computed. A PPP close to .5 would indicate consistency between the observed item association and the model prediction, and deviation in either direction would indicate misfit.

Figure 1.13 presents the PPPs of the ORs for the RLCM in the top-left corner and of the rRUM in the bottom-right. The vertical and horizontal axes represent the item indices, and the PPP for every item pair is represented by the shaded areas of the corresponding circle. The number of circles with close to 1/2 of shaded area is observed to be larger for the RLCM than for the rRUM. Whereas 28.31% of the PPPs lied above .95 or below .05 for the rRUM, only 10.32% of the PPPs were in this extreme range for the RLCM. We thus conclude that the RLCM was observed to provide a better fit in terms of item pairwise relationships.

1.5 Discussion

In this chapter we introduced a Bayesian estimation method for the restricted latent class model satisfying Xu's (2017) sufficient conditions for parameter identifiability. Results from simulation and empirical studies suggested accurate parameter recovery, efficient computation, and improved fit to the ECPE data in comparison to the more parsimonious rRUM.

The RLCM is by far the most general cognitive diagnosis model satisfying parameter

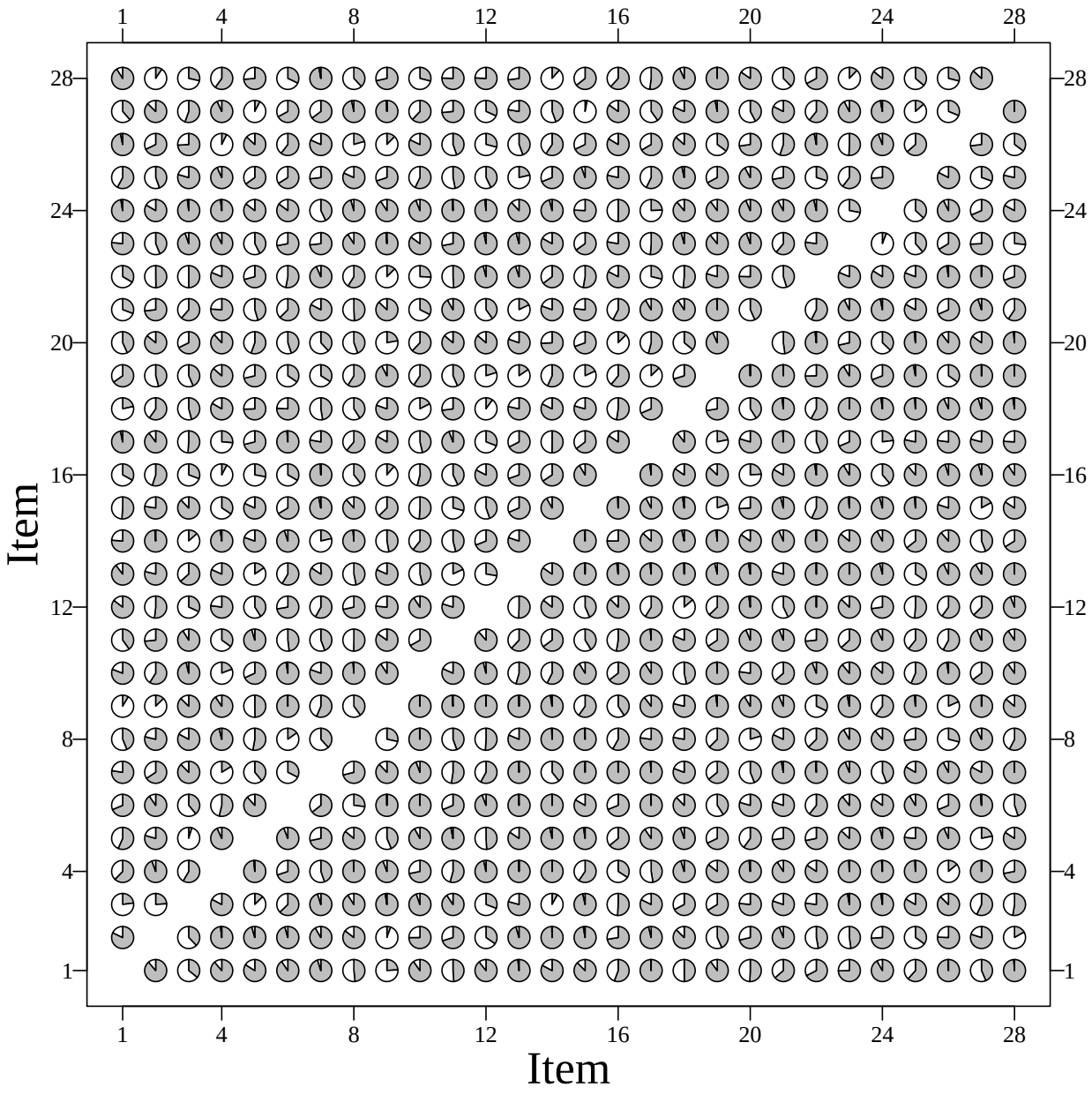

Figure 1.13: Posterior predictive probabilities of the observed item pairwise odds ratios for the RLCM (top-left) and the rRUM (bottom-right). Each bubble represents an item pair, and the shaded area on each bubble represents the posterior predictive probability of the observed odds ratio for that pair of item.

identifiability and subsumes many existing CDMs (e.g., Junker & Sijtsma, 2001; Henson & Templin, 2007; Templin & Henson, 2006; von Davier, 2008; de la Torre, 2011; Hartz, 2002). It relaxes two main assumptions that are usually imposed by other CDMs. First, it is typically assumed in CDMs that for any item and any two attribute patterns $\boldsymbol{\alpha}_1 \succeq \boldsymbol{\alpha}_2$, the correct

response probability of the latter cannot be higher than the former. A scenario in which this monotonicity assumption could be violated is when there are distractors: For example, suppose $K = 3$, and we consider a multiple choice item requiring all three attributes. To better differentiate examinees who have mastered all skills versus only the first two skills, one of the distractors of the item may specifically attract individuals who possess the first two skills. In this case, it is possible that individuals with $\boldsymbol{\alpha} = [1, 0, 0]$ may have higher correct response probability than those with $\boldsymbol{\alpha} = [1, 1, 0]$, because the latter is more prone to select the incorrect distractor. Under the RLCM, however, two nonzero patterns, both lacking 1 or more required skills for an item, can have correct response probabilities of any relationship.

Second, the other typical CDM assumption that is not imposed by the RLCM is the effect of non-required attributes on correct response probability. Whereas most of the general CDMs require the item response probability to be independent from the absence or presence of attributes not required by the Q-matrix, this independence assumption could be violated if possession of a non-required skill can increase the chance of correct response using an alternative strategy. The RLCM allows two examinees with same pattern on the required skill but different patterns on the non-required skills to have different correct response probabilities, as long as they lack some required skills by the item and both correct response chances are less than people who mastered all skills.

One could imagine that for the ECPE application described above, without more complex modeling of distractors or alternative strategies, some items might be discarded due to lack of fit if more restricted CDMs, such as the rRUM, are used. However, the RLCM is more flexible than the rRUM and hence the item and attribute patterns could be more accurately recovered if the RLCM is used. On consequence is that poorly fitting items under the rRUM are retained when using the RLCM rather than being discarded.

On the other hand, the RLCM can be further restricted and reduced to more parsimonious CDMs. For instance, the DINA (Junker & Sijtsma, 2001) model can be seen as a

special case of the RLCM when the probability of correct response is the same for all $\boldsymbol{\alpha}_c \not\succeq \mathbf{q}_j$; the DINO (Templin & Henson, 2006) model requires probability of correct response to be the same for $\{\boldsymbol{\alpha}_c : \alpha_{ck} = 1, \text{some } k : Q_{jk} = 1\}$ and for $\{\boldsymbol{\alpha}_c : \alpha_{ck} = 0, \forall k : Q_{jk} = 1\}$; and the L-CDM model (Henson et al., 2009) imposes the monotonicity inequality constraint and the independence equality constraint described above. By aggregating over equivalence classes for equality constraints and truncating the prior distribution of the item parameters based on inequality constraints, the proposed Bayesian Gibbs sampling algorithm could readily be modified to fit more restricted CDMs that are subsumed by the RLCM. One possible direction for future research is to further extend the proposed Gibbs sampling algorithm, so that any additional equality or inequality constraints can be easily incorporated in the parameter estimation. The potential to fit more restricted models also provides the possibility of using the proposed Bayesian estimation routine for model comparison. Posterior predictive model checking and other Bayesian model testing approaches can be used to assess the additional gain of including extra parameters and to identify the best fitting model for each item. Future studies can look into the development of model comparison and model testing methods under the Bayesian framework for the RLCM.

Chapter 2

Assessing Learning with the Reduced RUM

2.1 Introduction

With the rapid development in information technology, we see a growing number of online learning or tutoring systems that can be used by students either in or outside the classrooms. To enable real-time tracking of learners' progress, assessment are transforming from long, end-of-the-course tests to a trajectory of small quizzes embedded in the instruction (U.S. Department of Education, 2016). The response data collected from the assessments throughout the learning process allows researchers to develop models for learning, which can be used to (1) track students' progress over time, (2) evaluate the effectiveness of learning interventions, and (3) discover factors that may affect learning outcomes. Because students' acquisition of the contents cannot be directly observed and needs to be measured through assessment questions, psychometrics serves an important role in the learning models and provides a potent tool to assist learning (Chang, 2015; Zhang & Chang, 2016).

A lot of recent research considered the application of cognitive diagnosis models (CDMs) in the modeling of learning. Under these models, the learning process is considered as the repeated alternation between a learning stage, in which students are provided with materials that can improve their mastery of knowledge or skills, and an assessment stage, in which students are administered test questions to measure their mastery of skills in the curriculum at that time point. The students' progress throughout the learning process can be represented with changes in the attribute patterns over time. Markovian models are usually used for the transition between mastery and non-mastery on skills at each learning stage, and CDMs are

used to infer the unobserved mastery profiles based on the observed responses.

Previous research on learning models based on CDMs often adopted the deterministic input, noisy-"and"-gate (DINA) model as the measurement model. However, the DINA model's strict assumptions may lead to a lack of fit of most available items. The current chapter introduces learning models using two alternative measurement models, including the reduced-reparameterized unified model (rRUM, Hartz, 2002) and the noisy input, deterministic-"and"-gate (NIDA) model. We further consider two types of transition models: One is the simple independent monotonic transition model, and the other considers the existence of hierarchical relationships among attributes.

In the subsequent sections, we provide an incomplete overview of learning models based on CDMs, introduce the current models, propose a Bayesian modeling framework and an MCMC-based estimation algorithm, and evaluate the models' fit to a data set on the learning of spatial rotation skills.

2.1.1 Learning Models Based on CDMs

Whereas traditional research on CDMs focused on the assessment of the students' mastery at a single time point, there is increasing interest in assessing the students' progression of attribute mastery over time. Li, Cohen, Bottge, and Templin (2015) combined latent transition analysis (LTA; Langeheine, 1988; Collins & Wugalter, 1992) with the DINA model to estimate the students' mastery change in a longitudinal setting. At each wave, a student has a probability of transitioning from non-mastery to mastery, or from mastery to non-mastery, on each skill. The transitions on different skills are assumed to be independent. Using the estimated transition probabilities between mastery and non-mastery on each skill between waves, they compared the effectiveness of two different learning interventions.

Chen, Culpepper, Wang, and Douglas (2017) generalized Li et al.'s (2015) model to the case where attribute transitions are not necessarily independent — in other words, instead of modeling the attribute-wise transitions and multiplying them to get the pattern-wise

transition probabilities, they directly modelled the transition probabilities between different skill patterns. Different models for learning, including the most general unrestricted model and the monotonic model (i.e., where the probability of transitioning from mastery to non-mastery is 0), were introduced, and the cardinalities of the sets of all possible trajectories were derived.

S. Wang et al. (2016) proposed the higher-order hidden Markov cognitive dianosis model (HO-HM CDM) with higher-order covariates affecting the learning outcome. Given the attribute pattern of subject i at time t, $\boldsymbol{\alpha}_{i,t} = [\alpha_{i,1,t}, \ldots, \alpha_{i,K,t}]'$, the logit of the probability of transitioning from non-master to master on attribute k is

$$\text{logit}[P(\alpha_{i,k,t+1} = 1 \mid \alpha_{i,k,t} = 0, \boldsymbol{\alpha}_{i,t})] = \lambda_0 + \lambda_1\theta_i + \lambda_2 \sum_{\forall k' \neq k} \alpha_{i,k',t} + \lambda_3 \sum_{m=1}^{t} \sum_{j=1}^{J_t} q_{j,m,k}. \quad (2.1)$$

In this model, θ_i was used to denote the overall, time-invariant learning ability of subject i. The term $\sum_{\forall k' \neq k} a_{i,k',t}$ represents how many attributes subject i has already acquired other than attribute k, and $\sum_{m=1}^{t} \sum_{j=1}^{J_t} q_{j,m,k}$ denotes the number of items involving skill k that the student has completed at previous time points, in other words, the amount of practice so far on attribute k. By using a higher order logistic model for the transition probabilities in the hidden Markov model, the effect of different factors on the probability of learning a skill can hence be examined.

2.2 Current Model

Our proposed model can be regarded as the combination of two parts, a learning model that describes the transition of attribute patterns over time, and a measurement model that assesses the learners' skill mastery at each time point. The section below provides an overview for these two components.

2.2.1 Learning Model

We denote the attribute pattern for subject $i \in \{1, \ldots, N\}$ at time $t \in \{1, \ldots, T\}$ by $\boldsymbol{\alpha}_{i,t} = (\alpha_{i,t,1}, \ldots, \alpha_{i,t,K})'$. Here, $t = 1$ represents the initial time point before any learning takes place, and $t = 2, \ldots, T$ represent each subsequent time point. Two transition models are considered below, a general monotonic Markov model and a restricted model with attribute hierarchies.

Monotonic Markov model. Given the attribute pattern of subject i at time t, $\boldsymbol{\alpha}_{i,t} = (\alpha_{i,t,1}, \ldots, \alpha_{i,t,K})'$, the probability that the student masters skill $k \in \{1, \ldots, K\}$ at time $t + 1$ is given by

$$P(\alpha_{i,t+1,k} = 1 \mid \boldsymbol{\alpha}_{i,t}, \tau_k) = \begin{cases} 1, & \alpha_{i,t,k} = 1 \\ \tau_k, & \alpha_{i,t,k} = 0 \end{cases}, \tag{2.2}$$

where τ_k is the probability of transitioning from non-master to master on skill k at any given time point. If we denote the observed mastery status on attribute k at time $t + 1$ by a we have

$$P(\alpha_{i,t+1,k} = a \mid \boldsymbol{\alpha}_{i,t,k}, \tau_k) = P(\alpha_{i,t+1,k} = 1 \mid \boldsymbol{\alpha}_{i,t}, \tau_k)^a [1 - P(\alpha_{i,t+1,k} = 1 \mid \boldsymbol{\alpha}_{i,t}, \tau_k)]^{1-a}. \tag{2.3}$$

The probability of attribute pattern $\boldsymbol{\alpha}_{i,t+1}$, given $\boldsymbol{\alpha}_{i,t}$ and $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_K)'$, is

$$P(\boldsymbol{\alpha}_{i,t+1} \mid \boldsymbol{\alpha}_{i,t}, \boldsymbol{\tau}) = \prod_{k=1}^{K} P(\alpha_{i,t+1,k} \mid \alpha_{i,t,k}, \tau_k). \tag{2.4}$$

Restricted model with attribute hierarchies. The monotonic Markov model above assumes that the probability of learning an attribute does not depend on whether another attribute is also learned. In practice, however, some attributes may be prerequisite to others, and as a result, we may have a restricted set of possible attribute patterns. This will limit the number of possible attribute patterns from all 2^K possibilities. Denote the set of prerequisites

to attribute k as $\{\bar{k}\}$, then instead of equation (1), the probability of the transition is given by

$$P(\alpha_{i,t+1,k} = 1 \mid \boldsymbol{\alpha}_{i,t}, \tau_k) = \begin{cases} \prod_{k' \in \{\bar{k}\}} \alpha_{i,t+1,k'}, & \alpha_{i,t,k} = 1; \\ \tau_k \cdot \prod_{k' \in \{\bar{k}\}} \alpha_{i,t+1,k'}, & \alpha_{i,t,k} = 0. \end{cases} \tag{2.5}$$

The hierarchical relationship between attributes can be captured by using a $K \times K$ reachability matrix, $\boldsymbol{R}$ (e.g., Leighton, Gierl, & Hunka, 2004), where $R_{kk'} = 1$ if attribute k requires attribute k' as a prerequisite, and $R_{kk'} = 0$ otherwise.

2.2.2 Measurement Models

Here we consider two possible response models, the NIDA model (Junker & Sijtsma, 2001) and the rRUM model (Hartz, 2002). At a certain time point, we index the items by $j = 1, \ldots, J_t$, where J_t is the number of items administered at time t. Under the NIDA model, the probability of a correct response is given by

$$P(X_{i,t,j} = 1 \mid \boldsymbol{\alpha}_{i,t}, \mathbf{s}, \mathbf{g}) = \prod_{k=1}^{K} [(1 - s_k)^{\alpha_{i,t,k}} g_k^{(1-\alpha_{i,t,k})}]^{q_{j,k}}, \tag{2.6}$$

where $\mathbf{s} = [s_1, \ldots, s_K], \mathbf{g} = [g_1, \ldots, g_K]$ can be interpreted as the probabilities of incorrectly applying an acquired attribute (slipping) and probabilities of correctly applying an unaquired attribute (guessing), respectively.

By relaxing the slipping and guessing parameters to be item-specific $(\mathbf{s}_j, \mathbf{g}_j)$, we have the Generalized NIDA (GNIDA) model, given by

$$P(X_{i,t,j} = 1 \mid \boldsymbol{\alpha}_{i,t}, \mathbf{s}_j, \mathbf{g}_j) = \prod_{k=1}^{K} [(1 - s_{j,k})^{\alpha_{i,t,k}} g_{j,k}^{(1-\alpha_{i,t,k})}]^{q_{j,k}}. \tag{2.7}$$

With a reparameterization of the GNIDA model through the following conversions,

$$\pi_j^* = \prod_{k=1}^{K}(1-s_{j,k})^{q_{j,k}}, \tag{2.8}$$

$$r_{j,k}^* = \frac{g_{j,k}}{1-s_{j,k}}, \tag{2.9}$$

we obtain the rRUM model (Hartz, 2002), where the probability of correct response is given by

$$P(X_{i,t,j}=1 \mid \boldsymbol{\alpha}_{i,t}, \mathbf{q}_j, \pi_j^*, \mathbf{r}_j) = \pi_j^* \prod_{k=1}^{K} r_{j,k}^{*(1-\alpha_{i,t,k})q_{j,k}}. \tag{2.10}$$

2.3 Parameter Estimation

A Bayesian formulation is adopted to estimate the learning model's parameters. Similar to Culpepper and Hudson (2017), a data augmentation approach is used for updating the generalized NIDA (NIDA and rRUM) model parameters, and similar to S. Wang et al. (2016) and Chen, Culpepper, Shiyu, and Douglas (2017), the forward-backward algorithm was used for sequentially updating the learning model and the attribute parameters under the hidden Markov model. Under the current Bayesian formulation, the full conditional distributions of the parameters, given the data, are known families of distributions and thus can be directly sampled from using a Gibbs sampling algorithm.

2.3.1 Prior distribution

We assume that the prior distribution for the initial population membership probabilities, $\boldsymbol{\Pi}$, is

$$\boldsymbol{\Pi} \sim \text{Dirichlet}(\boldsymbol{\delta}_0), \text{where } \boldsymbol{\delta}_0 = (\delta_{01}, \ldots, \delta_{0C})', \text{ with } C = 2^K. \tag{2.11}$$

We further assume the prior distributions for the transition probabilities $\boldsymbol{T}$, are

$$p(\boldsymbol{T} = \boldsymbol{\tau}) \propto \prod_{k=1}^{K} \tau_k^{a-1}(1-\tau_k)^{b-1}. \tag{2.12}$$

In addition, for both the NIDA and the rRUM model truncated Beta priors were used for $s_{j,k}$s and $g_{j,k}$s, the slipping and guessing parameters under the G-NIDA formulation (note that under the NIDA model, the $s_{j,k}$s and $g_{j,k}$s were constrained to be equal across items and hence could be simplified to s_k and g_k):

$$p(s_{j,k}, g_{j,k}) \propto s_{j,k}^{a_s-1}(1-s_{j,k})^{b_s-1} g_{j,k}^{a_g-1}(1-g_{j,k})^{b_g-1}\mathcal{I}(0 \leq g_{j,k} < 1 - s_{j,k} \leq 1). \tag{2.13}$$

2.3.2 Full conditional distributions

Let $\mathbf{Z}_{i,t,k,\cdot} = (Z_{i,t,j,1}, \ldots, Z_{i,t,j,K})'$ denote the augmented latent responses to item j by subject i at time t, where $Z_{i,t,j,k} = 1$ if subject i has successfully applied skill k on item j at time t, and $Z_{i,t,j,k} = 0$ otherwise. In addition, let $\mathbf{Z}_{i,t,j,(k)}$ denote the vector of the latent responses on item j by subject i at time t except on attribute k. With the assumed prior distributions of the parameters described above, the full conditional distributions for the parameters, given the observed responses $x_{i,t,j}$s, are described below.

- For $Z_{i,t,j,k}$s such that the corresponding $q_{j,k} = 1$, the conditional distribution given the data and other parameters is the same as in Culpepper and Hudson (2017):

$$Z_{i,t,j,k} \mid (X_{i,t,j} = x_{i,t,j}, \mathbf{Z}_{i,t,j,(k)}, \alpha_{i,t,k}, s_{j,k}, g_{j,k}) \sim \text{Bernoulli}(\tilde{\pi}_{i,t,j,k}), \text{where} \tag{2.14}$$

$$\begin{aligned} \tilde{\pi}_{i,t,j,k} &= \frac{P(x_{i,t,j} \mid \mathbf{Z}_{i,t,j,(k)}, Z_{i,t,j,k} = 1) P(Z_{i,t,j,k} = 1 \mid \alpha_{i,t,k}, s_{j,k}, g_{j,k})}{\sum_{z_{i,t,j,k}=0}^{1} P(x_{i,t,j} \mid \mathbf{Z}_{i,t,j,(k)}, Z_{i,t,j,k} = z_{i,t,j,k}) P(Z_{i,t,j,k} = z_{i,t,j,k} \mid \alpha_{i,t,k}, s_{j,k}, g_{j,k})} \\ &= \{(1 - \prod_{k' \neq k} z_{i,t,j,k'}^{q_{j,k'}})[(1-s_{j,k})^{\alpha_{i,t,k}} g_{j,k}^{1-\alpha_{i,t,k}}]^{q_{j,k}}\}^{1-x_{i,t,j}}. \end{aligned} \tag{2.15}$$

- For $\alpha_{i,t,k}$'s: Let $\mathbf{s}_{\cdot k}, \mathbf{g}_{\cdot k}$ denote the vectors of slipping and guessing parameters associated with applying skill k for all items administered to subject n at time t, and let $\boldsymbol{\alpha}^*$ denote the attribute vector of length K, whose kth entry is $\alpha_{i,t,k}$ and the other entries are equal to $\boldsymbol{\alpha}_{i,t,(k)}$. Then

$$\begin{aligned} p(\alpha_{i,t,k} \mid \mathbf{z}_{i,t,\cdot,k}, \boldsymbol{s}_{\cdot,k}, \boldsymbol{g}_{\cdot,k}, \boldsymbol{\alpha}_{i,t,(k)}) &\propto p(\mathbf{z}_{i,t,\cdot,k} \mid \alpha_{i,t,k}, \boldsymbol{s}_{\cdot,k}, \boldsymbol{g}_{\cdot,k})\tilde{\pi}_{i,t,k} \\ &\propto [\prod_{j=1}^{J_t} P(z_{i,t,j,k} \mid \alpha_{i,t,k}, s_{j,k}, g_{j,k})]\tilde{\pi}_{i,t,k}, \quad (2.16) \end{aligned}$$

where

$$\tilde{\pi}_{i,t,k} = \begin{cases} P(\boldsymbol{\alpha}_{i,t} = \boldsymbol{\alpha}^* \mid \boldsymbol{\pi})P(\boldsymbol{\alpha}_{i,t+1} \mid \boldsymbol{\alpha}_{i,t} = \boldsymbol{\alpha}^*, \boldsymbol{\tau}), & t = 1; \\ P(\boldsymbol{\alpha}_{i,t} = \boldsymbol{\alpha}^* \mid \boldsymbol{\alpha}_{i,t-1}, \boldsymbol{\tau})P(\boldsymbol{\alpha}_{i,t+1} \mid \boldsymbol{\alpha}_{i,t} = \boldsymbol{\alpha}^*, \boldsymbol{\tau}), & 1 < t < T; \\ P(\boldsymbol{\alpha}_{i,t} = \boldsymbol{\alpha}^* \mid \boldsymbol{\alpha}_{i,t-1}, \boldsymbol{\tau}), & t = T, \text{ and} \end{cases} \quad (2.17)$$

$$P(z_{i,t,j,k} \mid \alpha_{i,t,k}, s_{j,k}, g_{j,k}) = [(1 - s_{j,k})^{\alpha_{i,t,k}} g_{j,k}^{(1-\alpha_{i,t,k})}]^{z_{i,t,j,k}} [s_{j,k}^{\alpha_{i,t,k}} (1 - g_{j,k})^{1-\alpha_{i,t,k}}]^{1-z_{i,t,j,k}}. \quad (2.18)$$

- For $\boldsymbol{\pi}$: Denote the attribute patterns of all subjects at time $t = 1$ by $\boldsymbol{\alpha}_{\cdot,1}$, then

$$\boldsymbol{\pi} \mid \boldsymbol{\alpha}_{\cdot,1} \sim Dirichlet(\boldsymbol{\delta}_1 + \tilde{\boldsymbol{\delta}}), \text{ where } \tilde{\boldsymbol{\delta}} = (\sum_{i=1}^{N} \mathcal{I}(\boldsymbol{\alpha}_{i,1} = \boldsymbol{\alpha}_1), \ldots, \sum_{i=1}^{N} \mathcal{I}(\boldsymbol{\alpha}_{i,1} = \boldsymbol{\alpha}_C))'. \quad (2.19)$$

- For $\boldsymbol{s}_k, \boldsymbol{g}_k$'s: Given an item j, let t_i^* denote the time at which item j was administered to subject i, and let $\boldsymbol{\alpha}_{\cdot,\cdot}$ represent the attribute patterns for all subjects across all time

points. Then

$$\begin{aligned} P(s_{j,k}, g_{j,k} \mid \mathbf{z}_{\cdot,\cdot,\cdot,\cdot}, \boldsymbol{\alpha}_{\cdot,\cdot}) &\propto s_{j,k}^{a_{s,j,k}-1}(1-s_{j,k})^{b_{s,j,k}-1} g_{j,k}^{a_{g,j,k}-1}(1-g_{j,k})^{b_{g,j,k}-1} \\ &\times \mathcal{I}(1 \leq g_{j,k} < 1 - s_{j,k} \leq 1), \end{aligned} \tag{2.20}$$

where, under the rRUM model,

$$a_{s,j,k} = \sum_{i=1}^{N} \alpha_{i,t_i^*,k}(1 - z_{i,t_i^*,j,k})q_{j,k} + a_s; \tag{2.21}$$

$$b_{s,j,k} = \sum_{i=1}^{N} \alpha_{i,t_i^*,k} z_{i,t_i^*,j,k} q_{j,k} + b_s; \tag{2.22}$$

$$a_{g,j,k} = \sum_{i=1}^{N} (1 - \alpha_{i,t_i^*,k}) z_{i,t_i^*,j,k} q_{j,k} + a_g; \tag{2.23}$$

$$b_{g,j,k} = \sum_{i=1}^{N} (1 - \alpha_{i,t_i^*,k})(1 - z_{i,t_i^*,j,k}) q_{j,k} + b_g. \tag{2.24}$$

Under the NIDA model, the $s_{j,k}$ and $g_{j,k}$s are the same across items, thus $a_{s,j,k}, b_{s,j,k}, a_{g,j,k}$, and $b_{g,j,k}$ can be simplified to $a_{s,k}, b_{s,k}, a_{g,k}$, and $b_{g,k}$, where

$$a_{s,k} = \sum_{i=1}^{N} \sum_{k=1}^{K_{t_i^*}} \alpha_{i,t_i^*,k}(1 - z_{i,t_i^*,j,k})q_{j,k} + a_s; \tag{2.25}$$

$$b_{s,k} = \sum_{i=1}^{N} \sum_{k=1}^{K_{t_i^*}} \alpha_{i,t_i^*,k} z_{i,t_i^*,k} q_{j,k} + b_s; \tag{2.26}$$

$$a_{g,d} = \sum_{i=1}^{N} \sum_{k=1}^{K_{t_i^*}} (1 - \alpha_{i,t_i^*,k}) z_{i,t_i^*,j,k} q_{j,k} + a_g; \tag{2.27}$$

$$b_{g,d} = \sum_{i=1}^{N} \sum_{k=1}^{K_{t_i^*}} (1 - \alpha_{i,t_i^*,k})(1 - z_{i,t_i^*,j,k}) q_{j,k} + b_g. \tag{2.28}$$

- For $\boldsymbol{\tau}$: Let $\boldsymbol{\alpha}_{\cdot,t}$ denote the attribute patterns for all subjects at time t, and let $\{\bar{k}\}$

denote the set of prerequisites to attribute k. Then

$$P(\boldsymbol{\tau} \mid \boldsymbol{\alpha}_{\cdot,t}, \boldsymbol{\alpha}_{\cdot,t+1}) \propto \prod_{k=1}^{K} \tau_k^{a_{\tau_k}-1}(1-\tau_k)^{b_{\tau_k}-1}, \text{ with} \tag{2.29}$$

$$a_{\tau_k} = \sum_{t=1}^{T-1}\sum_{i=1}^{N}\left\{(1-\alpha_{i,t,k})\alpha_{i,t+1,k}\prod_{k'\in\{\bar{k}\}}\alpha_{i,t+1,k'}\right\} + a; \tag{2.30}$$

$$b_{\tau k} = \sum_{t=1}^{T-1}\sum_{i=1}^{N}\left\{(1-\alpha_{i,t,k})(1-\alpha_{i,t+1,k})\prod_{k'\in\{\bar{k}\}}\alpha_{i,t+1,k'}\right\} + b. \tag{2.31}$$

2.3.3 A Gibbs Sampling Algorithm

Because the full conditional distributions of all parameters can be directly sampled from, we can use a Gibbs sampler to iteratively draw samples of the parameters from the full conditional distributions. More specifically, the parameters were updated following these steps:

(1) Assign initial values to all parameters, namely $\boldsymbol{\pi}^{[0]}, \boldsymbol{\alpha}^{[0]}, \boldsymbol{s}^{[0]}, \boldsymbol{g}^{[0]}, \boldsymbol{\tau}^{[0]}$, and $\boldsymbol{z}^{[0]}$.

(2) At each iteration r:

(a) For each i, j, t, and $k : q_{j,k} = 1$, draw $z_{i,t,j,k}^{[r+1]}$ based on equation (2.14), given $x_{i,t,j}, \mathbf{z}_{i,t,j,(k)}^{[r]}, \alpha_{i,t,k}^{[r]}, s_{j,k}^{[r]}$, and $g_{j,k}^{[r]}$;

(b) For each i, t, and k, draw $\alpha_{i,t,k}^{[r+1]}$ based on equation (2.16), given $\mathbf{z}_{i,t,\cdot,k}^{[r+1]}, \boldsymbol{s}_{\cdot,k}^{[r]}, \boldsymbol{g}_{\cdot,k}^{[r]}, \boldsymbol{\alpha}_{i,t-1}^{[r+1]}$, $\boldsymbol{\alpha}_{i,t+1}^{[r]}, \boldsymbol{\pi}^{[r]}$, and $\boldsymbol{\tau}^{[r]}$;

(c) Draw $\boldsymbol{\pi}^{[r+1]}$ based on equation (2.19), given $\boldsymbol{\alpha}_{\cdot,0}^{[r+1]}$;

(d) For the rRUM model, for each item k and attribute d, draw $g_{j,k}^{[r+1]}$) based on equations (2.23) and (2.24), given $\mathbf{z}_{\cdot,\cdot,j,k}^{[r+1]}, \boldsymbol{\alpha}_{\cdot,\cdot}^{[r+1]}$ and $s_{j,k}^{[r]}$, and draw $s_{j,k}^{[r+1]}$ based on equations (2.21) and (2.22) given $\mathbf{z}_{\cdot,\cdot,j,k}^{[r+1]}, \boldsymbol{\alpha}_{\cdot\cdot}^{[r+1]}$ and $g_{j,k}^{[r+1]}$. The corresponding $\pi_k^{*[r+1]}$ and $\boldsymbol{r}_k^{*[r+1]}$ can be obtained via algebraic transformations in equations (2.8) and (2.9).

For the NIDA model, for each attribute k, draw $g_k^{[r+1]}$ based on equations (2.27) and (2.28), given $\mathbf{z}_{\cdot,\cdot,\cdot,k}^{[r+1]}, \boldsymbol{\alpha}_{\cdot,\cdot}^{[r+1]}$, and $s_k^{[r]}$, and draw $s_k^{[r+1]}$ based on equations (2.25) and (2.26), given $\mathbf{z}_{\cdot,\cdot,\cdot,k}^{[r+1]}, \boldsymbol{\alpha}_{\cdot,\cdot}^{[r+1]}$, and $g_k^{[r+1]}$.

(e) For each k, sample $\tau_k^{[r+1]}$ from the posterior distribution in equation (2.29), given $\boldsymbol{\alpha}_{\cdot,\cdot}^{[r+1]}, [\tau_1, \ldots, \tau_{k-1}]'^{[r+1]}$, and $[\tau_{k+1}, \ldots, \tau_K]'^{[r+1]}$.

2.4 Application: A Spacial Reasoning Test with Learning Interventions

Simulation studies were conducted to evaluated the performance of the Gibbs sampler in terms of parameter recovery. Results indicated that all parameters were able to be estimated with small bias when the chains were long enough. We skip the results of the simulations here and move directly to the real data application.

2.4.1 Spatial Rotation Data

The Spatial Rotation Data has been used by several previous researchers to evaluate the performance of proposed learning models (e.g., S. Wang et al., 2016; Chen, Culpepper, Wang, & Douglas, 2017). A computer-based assessment and training software was developed to conduct a study of learning spatial rotation skills. Subjects were students recruited from the paid subject pool of the Department of Psychology at the University of Illinois at Urbana-Champaign. Each subject was asked to complete a series of 50 items on a computer-based assessment on spatial rotation ability. The assessment items were comprised of an extended version of the Purdue Spatial Visualization Test (PSVT, Yoon, 2011). The assessment consisted of 5 test blocks, each containing 10 questions. Following each test block, except the final one, was a learning intervention. Participants first answered the 10 questions in a test block, then they proceeded to a learning block with instructions. The instruction were either an interactive figure, where the 3-D object in the demonstration question could be

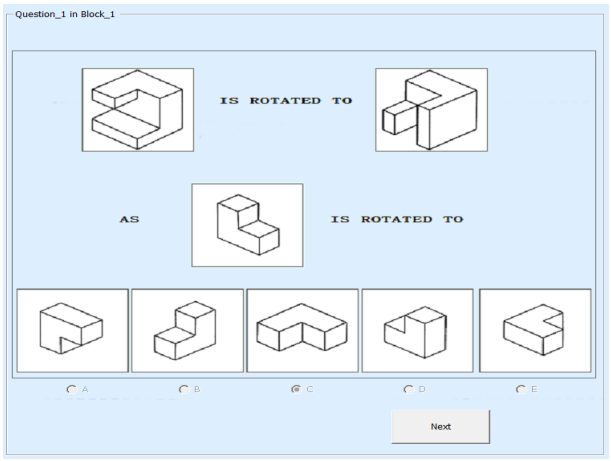

Figure 2.1: Test Block

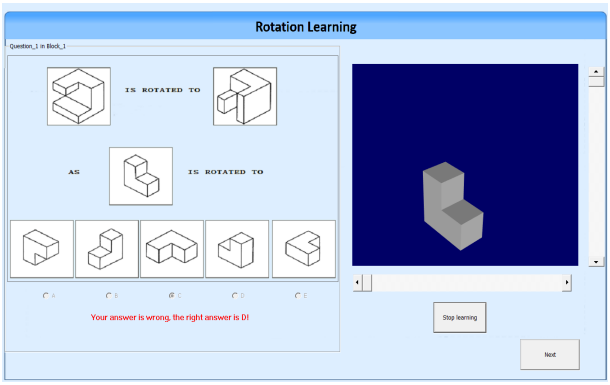

Figure 2.2: Learning Block: Type 1

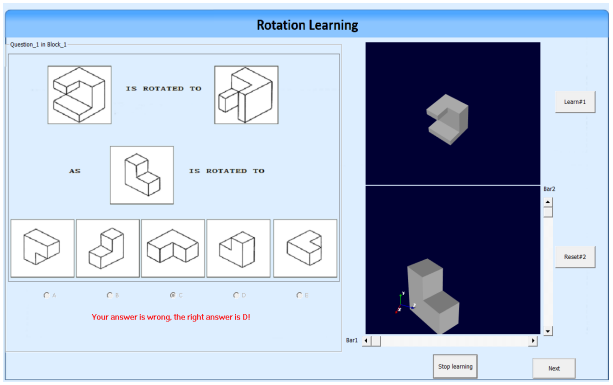

Figure 2.3: Learning Block: Type 2

moved freely along both the x-axis and the y-axis (Figure 2.2), or the interactive figure plus a recorded clip of the rotation of the object from the original position to the correct position (Figure 2.3).

Each item in an assessment block (Figure 2.1) featured a reference object that had undergone a rotation. Subjects then considered a new object and attempted to determine which of the 5 options corresponded to the same rotation as the reference object. All items included either an x-axis or y-axis rotation, or both, and varied in complexity. Four mental rotation skills were identified: 1) 90° x-axis rotation, 2) 90° y-axis rotation, 3)180° x-axis rotation and 4) 180° y-axis rotation.

Because we assumed the learning process to be monotonic, the proportion of students

mastering a large number of skills are expected to increase as the learning process continues. In that case, the proportion of students mastering few skills may be small at later time points. To ensure sufficient sample size for the estimation of the parameters of all items, the Spatial Rotation study used a block design for item assignment to subjects. Specifically, the N examinees were randomly assigned to 5 test design groups. For students in group 1, they were administered items in blocks 1, 2, 3, 4, and 5 at times $t = 1, 2, 3, 4, 5$, respectively. For students in group 2, they were administered item blocks $2, 3, 4, 5, 1$ at times 1 to 5. And students in group 3 were administered the item blocks in the order of $3, 4, 5, 1, 2$, and so on. A total of 351 University of Illinois students participated in this experiment.

2.4.2 Evaluated Models

Six different models, with two types of measurement models (NIDA or rRUM) and three types of attribute relationships, were compared in terms of fit to the Spatial Rotation data. The three different types of attribute relationships were captured by the corresponding reachability matrices. Specifically,

- Relationship 1: No attribute hierarchies exist, thus

$$R_1 = \begin{pmatrix} 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{pmatrix}. \tag{2.32}$$

- Relationship 2: 180° rotation along the y-axis requires 90° rotation along the y-axis as a prerequisite, and 180° rotation along the x-axis requires 90° rotation along the x-axis

as a prerequisite. The reachability matrix is hence

$$R_2 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}. \tag{2.33}$$

- Relationship 3: The 180° rotations (along x-axis and y-axis) has **both** the 90° rotation along x-axis and the 90° rotation along y-axis as prerequisites. The corresponding reacheability matrix is

$$R_3 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{pmatrix}. \tag{2.34}$$

The MCMC algorithm described previously was applied to estimate the model parameters. Uninformative priors were chosen for $\boldsymbol{\pi}, \boldsymbol{s}, \boldsymbol{g}$, and $\boldsymbol{\tau}$. The initial value of $\boldsymbol{\pi}$ was randomly sampled from Dirichlet($\mathbf{1}$), the initial values for each τ_k were sampled from the Uniform(0,1) distribution, and the initial values for the $s_{j,k}, g_{j,k}$'s were randomly sampled from $U(.1, .3)$. Using these random initial values, the initial values for $\boldsymbol{\alpha}$'s were simulated. Lastly, $z_{i,t,j,k}^{[0]}$ was set to 1 for all i, t, j and k.

2.4.3 Model convergence

To evaluate the model convergence, five separate chains with different starting values were run with chain lengths of $50,000$ iterations under the rRUM model with no attribute hierarchies. The Gelman-Rubin proportional scale reduction factor (PSRF), commonly known as $\hat{R}$, was calculated for each parameter at different chain lengths, and the progression of the maximum $\hat{R}$ out of all estimated parameters is displayed in Figure 2.4.

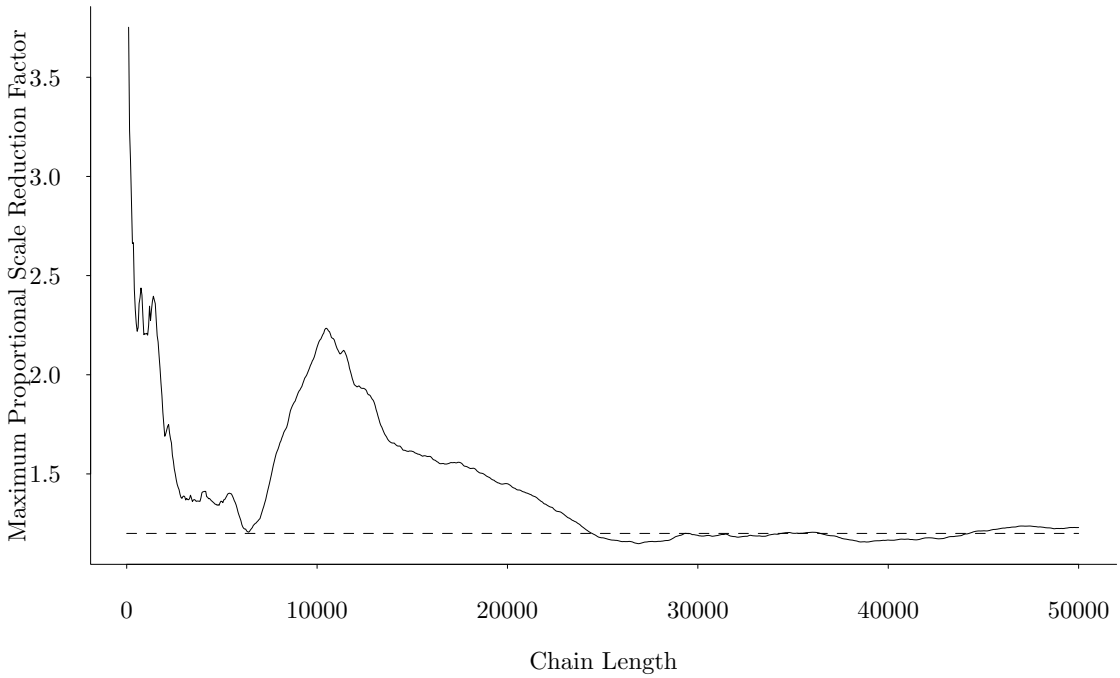

Figure 2.4: Progression of maximum $\hat{R}$ as chain length increases. Dashed horizontal line indicates $\hat{R} = 1.2$.

An $\hat{R}$ value of below 1.2 is commonly used in the literature as an indicator of the convergence of that parameter estimate. We can see that at around 25000 iterations, the maximum $\hat{R}$ stabilizes to less than or slightly above (up to .02 above) 1.2, and the $\hat{R}$ of all the other estimated parameters stay below 1.2. Thus for subsequent analyses, a chain length of 40000 was used for each of the 6 models, with 25000 iterations as the burn-in.

2.4.4 Model Comparison

The fit of the 6 models were compared in terms of a few aspects, the Deviance Information Criterion (DIC), and posterior predictive model checks on the item means (first moments, M1), item pair-wise odds ratios (second moments, M2), and on the subjects' total scores across time points. The procedures for computing each are detailed below.

1. DIC: As described in Spiegelhalter et al. (2002), if we denote the set of unknown model

parameters by θ, then the DIC can be calculated as

$$DIC = p_D + \bar{D}(\theta), \tag{2.35}$$

where $p_D = \bar{D}(\theta) - D(\bar{\theta})$, and

$$\begin{aligned} D(\theta) &= -2\log[P(\mathbf{x} \mid \theta)] \\ &= -2\log[P(\mathbf{x} \mid \boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{s}, \boldsymbol{g})] \\ &= -2\log\Big\{\prod_{i=1}^{N}\Big[\sum_{\forall \boldsymbol{\alpha}_l \in \mathcal{A}^T} P(\boldsymbol{\alpha}_l)\prod_{t=1}^{T} P(\boldsymbol{x}_{i,t} \mid \boldsymbol{\alpha}_{i,t} = \boldsymbol{\alpha}_{l,t}, \boldsymbol{s}_t, \boldsymbol{g}_t)\Big]\Big\}. \end{aligned} \tag{2.36}$$

Here, $\boldsymbol{\alpha}_l = (\boldsymbol{\alpha}_{l,1}, \ldots, \boldsymbol{\alpha}_{l,T})$ is any learning trajectory from time $t = 1$ to T, and $P(\boldsymbol{\alpha}_l)$ can be computed as

$$P(\boldsymbol{\alpha}_l) = P(\boldsymbol{\alpha}_{n,1} = \boldsymbol{\alpha}_{l,1} \mid \boldsymbol{\pi})\prod_{t=2}^{T} P(\boldsymbol{\alpha}_{i,t} = \boldsymbol{\alpha}_{l,t} \mid \boldsymbol{\alpha}_{i,t-1} = \boldsymbol{\alpha}_{l,t-1}, \boldsymbol{\tau}). \tag{2.37}$$

To calculate $\bar{k}(\theta)$, at each post-burnin iteration t of the MCMC, we compute the $D(\theta^{[t]})$ based on the parameter samples in the tth iteration, namely $\theta^{[t]}$. The average of $D(\theta^{[t]})$s across all post-burnin iterations is computed to obtain $\bar{k}(\theta)$.

2. Posterior predictive check for the item means (M1): After the burnin, at each iteration of the MCMC, the model parameter samples were used to simulate responses of the subjects. For each item, the proportion of people who answered correctly was calculated based on the simulated responses as well as the observed data. Then the posterior predictive probability (PPP) of each item's mean is given by the proportion of simulated item means that lie below the observed item mean.

3. Posterior predictive check for the item pairwise odds ratios (M2): For any given item pair, Sinharay (2006) suggested using $OR = (n_{11}n_{00})/(n_{01}n_{10})$, where n_{11} is number of respondents responding to both items correctly, n_{01} is number of respondents who

answered item 1 wrong and item 2 correctly, etc., as a measure of item-pairwise associations. Similar to that of the item means, the item pair-wise ORs based on the simulated responses from sampled model parameters and those from the observed responses are obtained, and the PPP of each item pair's odds ratio is given by the proportion of simulated odds ratios for the item pair that lie below the observed.

4. Posterior predictive model check for the subjects' total scores at each time point: Like above, simulated and observed responses were used to obtain the number of correct responses (total score) by each subject at each time point. Then, for each subject and each time point, the PPP for total score is given by the proportion of simulated total scores below the observed.

Table 2.4.4 summarizes the DIC statistics and the proportions of posterior predictive probabilities below .05 or above .95 (which indicates misfit) for item means (M1), item ORs (M2), and subject total scores for the six models. A smaller DIC value and a smaller proportion of PPPs outside the 90% interval would indicate better fit.

Table 2.1: Summary of fit statistics of the six different models. M1 misfit represents the percentage of item means that were outside the 95% posterior prediction interval, M2 misfit represnets percentage of item pair-wise odds ratios outside the prediction interval, and total misfit represents the percentage of observed total scores at each time point outside the 95% posterior prediction interval.

Model	DIC	M1 misfit	M2 misfit	total misfit
NIDA R_1	16129.61	74%	46.4%	24.1%
NIDA R_2	16154.09	70%	47.1%	23.1%
NIDA R_3	16233.26	72%	48.2%	23.1%
rRUM R_1	14860.91	0%	25.1%	23.5%
rRUM R_2	15099.09	0%	26.4%	23.6%
rRUM R_3	15188.04	0%	27.3%	23.8%

Table 2.4.4 suggests that out of the six models, the one assuming a rRUM measurement model and no attribute hierarchies achieved the best fit, indicated by the lowest DIC, the lowest proportion of extreme PPPs on item means and pair-wise odds ratios, and comparable proportion of extreme PPPs as the other models.

Figure 2.5 presents the posterior predictive probabilities of each item's mean under the rRUM model without attribute hierarchies. The shaded area in each circle represents the proportion of simulated item means below the observed item mean. None of the observed item means were within the extreme range. There is a consistent tendency for the model to slightly underestimate the item means, as indicated by the PPPs above 50% on all items.

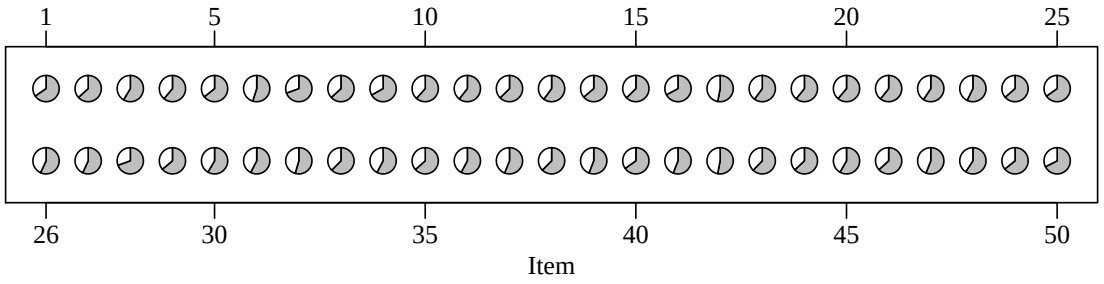

Figure 2.5: Posterior Predictive Probabilities (PPPs) of the item means (i.e., proportion correct).

Figure 2.6 presents the density of the posterior predictive probability of the item pair-wise odds ratios. We observe that the distribution of the PPPs is skewed to the left, indicating a tendency for the model to underestimate the ratio $(n_{11}n_{00})/(n_{01}n_{10})$.

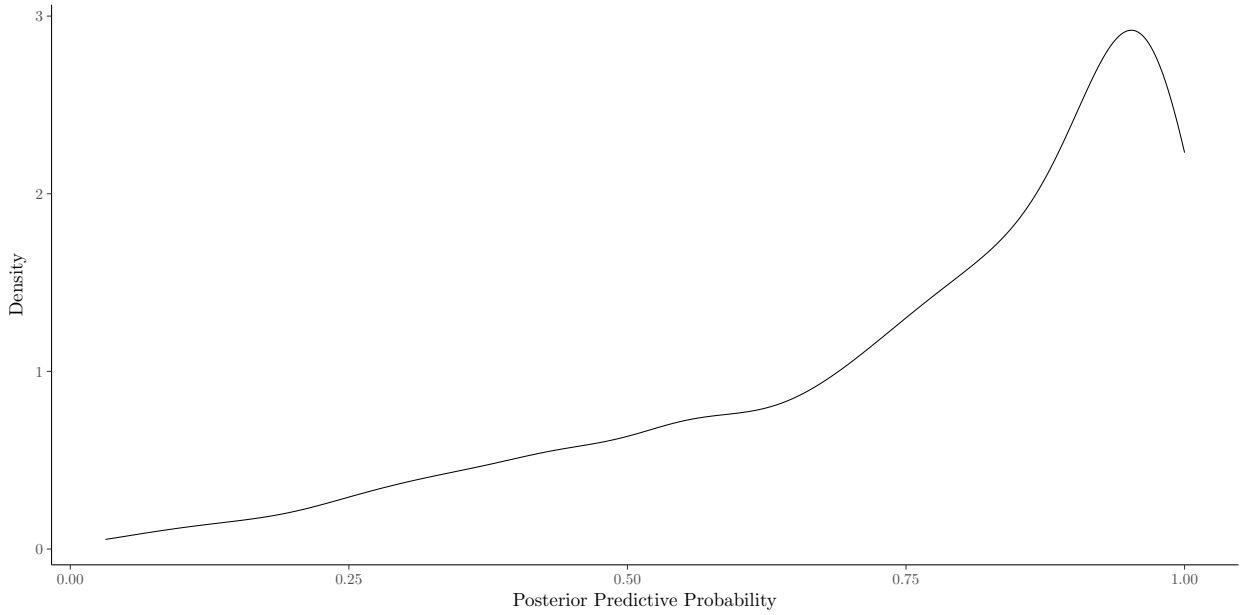

Figure 2.6: Density of the posterior predictive probability for item pair-wise odds ratios.

Figure 2.7 presents the density curves of the posterior predictive probabilities for total scores at different time points. For all time points, we observe a tendency for the model to underestimate the total scores of the subjects, as suggested by the higher densities at higher PPPs. This pattern seems to be most salient after the learning begins (i.e., for $T = 2, \ldots, 5$) than for the initial time point, $T = 1$.

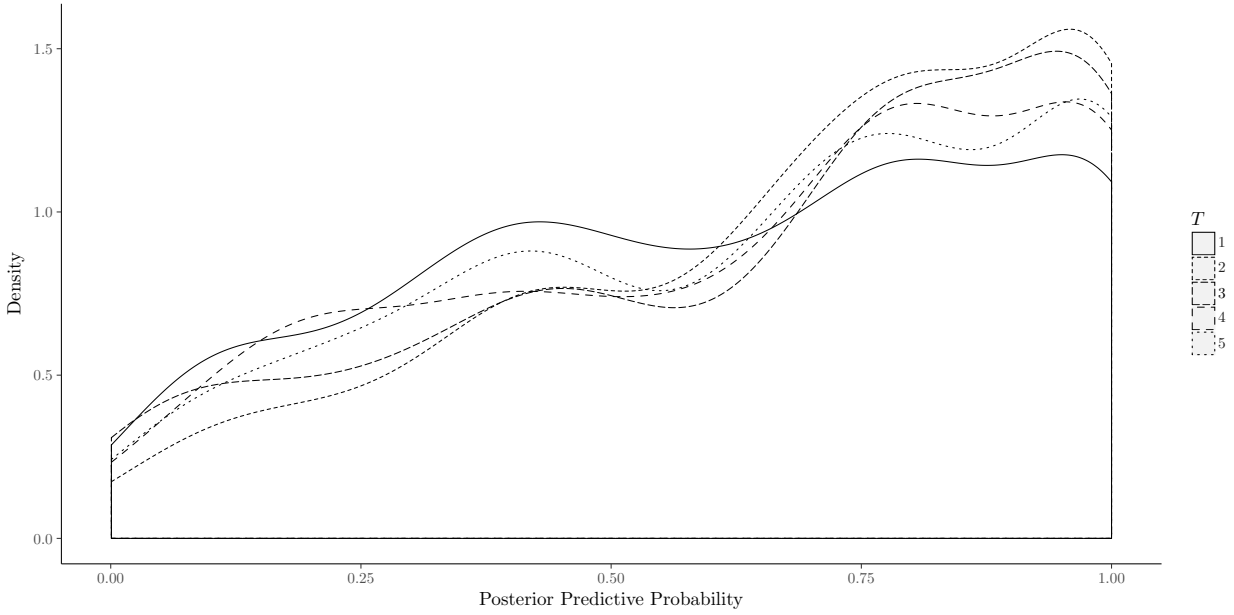

Figure 2.7: Density of posterior predictive probability for total scores at different time points.

The plots for PPPs of total scores over observed total score at each time point are shown in Figure 2.8. Across all time points, we observe a consistent trend for the model to overestimate total scores for subjects with low observed scores and underestimate for those with high observed scores.

2.4.5 Observed progression of learning

Based on the estimated attribute patterns under the rRUM learning model without attribute hierarchies, we looked at the progression of skill mastery rate over time, as well as the frequency for the number of mastered skills at each time point. Table 2.2 summarizes the distribution of the number of mastered skills at each time point, and Figure 2.9 shows the

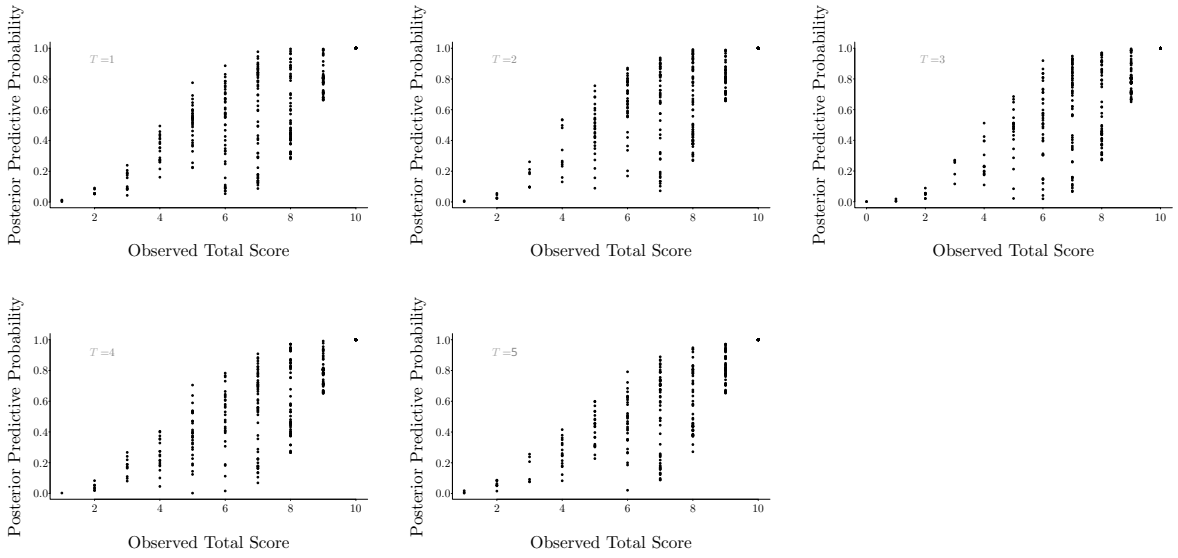

Figure 2.8: Relationship between posterior predictive probability of total score and observed total score at each time point.

progression of mastery rate of each skill across time. As the learning time increases, the percentage of students mastering each skill also increases, and a shift towards mastering more skills over time is observed.

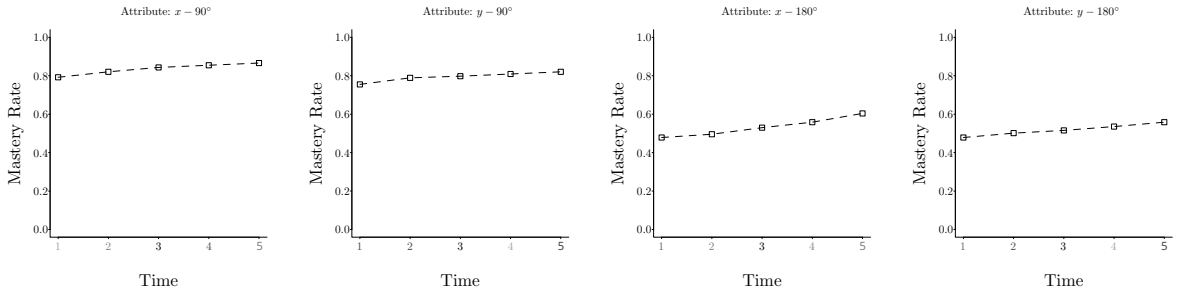

Figure 2.9: Progression of mastery rate of each skill across time.

Table 2.2: Frequency (and percentage) of number of skills mastered over time.

Skills mastered	T =1	T =2	T =3	T =4	T =5
0	53(15.1%)	40(11.4%)	35(9.97%)	31(8.83%)	27(7.69%)
1	53(15.1%)	57(16.24%)	56(15.95%)	56(15.95%)	56(15.95%)
2	77(21.94%)	76(21.65%)	70(19.94%)	58(16.52%)	45(12.82%)
3	0(0.00%)	6(1.71%)	13(3.7%)	28(7.98%)	38(10.83%)
4	168(47.86%)	172(49%)	177(50.43%)	178(50.71%)	185(52.71%)

2.5 Discussion

In the current chapter, we presented a learning model that combines the rRUM measurement model with a simple Markov model for learning, in which the transitions of separate skills were considered independent. The rRUM model without attribute hierarchies provided a reasonably good but imperfect fit to the spatial rotation data. In particular, posterior predictive checks found the predicted total scores over time did not advance like observed scores. This is likely due to the simplicity of the learning model by treating attributes as independent. On the other hand, the posterior predictive model checks of the item parameters suggested that the rRUM can provide a relatively good fit for the assessment items, at least approximating the first moments very well.

Future research can combine the rRUM with more complex transition models allowing dependence between skills, such as the hidden Markov Diagnostic Classification Model (S. Wang et al., 2016) and the first order hidden Markov model with transitions of skill patterns instead of single skills (Chen, Culpepper, Shiyu, & Douglas, 2017). The application also considered the notion of attribute hierarchies, which when present could greatly reduce the number of possible attribute patterns and speed up computations. However, in this application goodness of fit measures indicated the superior fitness of the unrestricted model.

Chapter 3

Mixture Learning Model with Responses and Response Times

3.1 Introduction

Educational researchers have shown long term interests in understanding the heterogeneity among online learners. Learners can differ not only in their initial background and general learning ability, but also in terms of how they learn. For example, learners' affects can influence their behaviors in the learning process and hence the learning outcomes. Methods were proposed by educational data miners to detect student affect based on their interactions with the online learning systems (e.g., Baker et al., 2012). By identifying the affects of each student during the learning process, such as boredom, disengagement, confusion, and frustration, educators can provide targeted interventions accordingly to improve learning outcomes. Students can also vary in their preferred mode of instructions. Felder and Silverman (1988) developed the Index of Learning Styles survey, which measured learners' characteristics on the Sensing/Intuiting, Visual/Verbal, Active/Reflective, and Sequential/Global dimensions. A student's learning style can provide indications of possible strengths and difficulties in the learning process.

The increased popularity of computer-based testing has enabled researchers to collect various types of process data, including test takers' reaction time to assessment items, also known as response times. In the field of psychometrics, extensive research has been conducted on the joint modeling of response accuracy and response times, which can improve the estimation accuracy of item parameters and examinees' latent traits or latent classes, further our understanding of individuals' test-taking behavior and the test items' characteris-

tics, and help us differentiate examinees using different test-taking strategies. Most recently, a higher-order hidden Markov CDM (HO-HM CDM) framework, which simultaneously accounts for changes in response accuracy and response times throughout the learning process, was developed to measure students' improvements in skill mastery over time.

Response times can also provide a rich source of information for identifying students' learning styles, especially student engagement. Henrie, Halverson, and Graham (2015) provided a comprehensive review of methods for measuring student engagement in technology-based learning environment in the literature, and the time spent on homework, web pages, readings, et cetera were commonly used as an indicator of student engagement. Response times were also used by educational data miners to identify disengaged learners (Beck, 2004). In the current chapter, we incorporate response times information into the modeling of learning styles under the hidden Markov learning model framework. Based upon the joint modeling framework under the HO-HM CDM, we propose a mixture model for learning with response times and response accuracy. On top of modeling learning trajectories and item responses, response time data, along with response data, are additionally used to identify subpopulations with different learning and test-taking behaviors. Such a model accounts for the presence of heterogeneities in learning styles among students and may provide instructors with valuable information, which can be used to design individualized instructions.

This chapter is organized as follows. We first describe the joint modeling framework of response times and response accuracy under the HO-HM CDM (S. Wang, Zhang, Douglas, & Culpepper, 2018). Then, we give an incomplete overview of mixture models in psychometrics research and mixture hidden Markov models. The mixture learning model with response times and response accuracy is then introduced. A Bayesian parameter estimation framework is proposed, followed by simulation studies verifying the parameter recovery. Lastly, we discuss some implications of the proposed model as well as future directions.

3.1.1 Joint Learning Model with Response Times and Response Accuracy

S. Wang et al. (2018) extended the original HO-HM CDM in Equation (2.1) to a joint modeling framework that incorporates both responses and response times on the assessment items. In addition to the original two components, namely the transition model and the measurement model for response accuracy, the joint modeling framework introduces a third component to the hidden Markov learning model, namely the measurement model for response times. The reaction times of the learners on each time is treated as another source of observed data that can be used to measure the latent attribute patterns ($\boldsymbol{\alpha}$s) and the latent speeds of learners. Specifically, the probability density of a learner i's reaction time (latency) to an item $j \in \{1, \ldots, J_t\}$ at time $t \in \{1, \ldots, T\}$, $L_{i,j,t}$, is given by

$$f(L_{i,j,t}|\tau_i, \gamma_j, a_j, \boldsymbol{\alpha}_{i,t}) = \frac{a_j}{L_{i,j,t}\sqrt{2\pi}} \exp(-\frac{1}{2}[a_j(\log(L_{i,j,t}) - (\gamma_j - \tau_i - \phi * G_{i,j,t}]^2). \tag{3.1}$$

In other words, $L_{i,j,t}$ follows a log-normal distribution,

$$\log(L_{i,j,t}) \sim N\left(\gamma_j - (\tau_i + \phi * G_{i,j,t}), \frac{1}{a_j^2}\right). \tag{3.2}$$

Here, τ_i is the initial latent speed of learner i, γ_j is the time-intensity parameter of item j, capturing the overall amount of time the item requires, and a_j is the time-discrimination parameter of item j, which captures variance of response times at given τ_i and γ_j. Unlike response time models for static assessments (e.g., van der Linden, 2006), which typically treated the latent speed as unchanged throughout the test, in the learning context, it is reasonable to assume that the speed of the learners increases over time. The increase in speed in responding to the items is captured in equation (3.1) by $G_{i,j,t}$. Depending on how response speed changes over time, G can either be independent or dependent on the latent trajectory $\boldsymbol{\alpha}_i$. S. Wang et al. (2018) proposed several possible versions of G, including

- $G_{i,j,t} = t_{i,j}/T$, where $t_{i,j}$ is the time point at which item j is assigned to subject i. This essentially assumes that the learners' latent speed increases at a constant rate over time, which is independent of the attribute trajectories $\boldsymbol{\alpha}_i$.

- $G_{i,j,t} = \begin{cases} 1, & \text{if } \boldsymbol{\alpha}_{i,t} \succeq \mathbf{q}_j, \\ 0, & \text{otherwise.} \end{cases}$
  This version of G assumes that the latent speed on an item increases by a constant when the learner has mastered all requisite skills of the item by the time the item is assigned. With this version of G, the response times depend on the current attribute pattern $\boldsymbol{\alpha}_{i,t}$, but is unaffected by the overall trajectory.

- $G_{i,j,t} = \log(\sum_{m<t}\sum_h \eta^*_{i,h,m} + \sum_{q<j}\eta^*_{i,q,t} + 1)$, where $\eta^*_{i,h,m} = (\prod_{k=1}^K \alpha_{i,m,k}^{q_{h,k}}) \cdot \mathcal{I}(\mathbf{q}_h^T\mathbf{q}_j > 0)$ indicates whether subject i has answered item h as a master at time m and whether items h and j have overlaps in requisite skills. Intuitively, $G_{i,j,t}$ counts the total number of questions relevant to the current item j on which the learner has answered as a master. It assumes that practices on mastered skills have a cumulative effect on the increase in latent speed. Using this version of G, it is assumed that the person's latent speed on an item depends on the entire trajectory of $\boldsymbol{\alpha}_i$ prior to this item.

Lastly, the slope in front of G, ϕ, captures the rate at which the learner's latent speed changes according to the covariate G. We note that, when $\phi = 0$, Wang's model reduces to the traditional log-normal response time model (van der Linden, 2006), where the distribution of response times is governed by the subject's latent speed and the item's time-intensity and time-discrimination.

The paragraphs above described the measurement model for response times. Under the joint modeling framework proposed by S. Wang et al. (2018), the DINA model in Equation (1.2) is used as the measurement model for response accuracy. On the structural level, how the learner's attribute pattern changes over time (i.e. the transition model, remains the same as in the HO-HM CDM in Equation (2.1)). The learner's initial speed, τ_i, can be

jointly modeled with his or her latent learning ability, θ_i, through a multivariate normal distribution, with $\begin{pmatrix} \theta_i \\ \tau_i \end{pmatrix} \sim MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. This joint modeling assumes a correlation between the subjects' initial latent speed and latent learning ability.

3.1.2 Mixture Models and Mixture Hidden Markov Models

Mixture Models. Mixture models are commonly used in statistics to describe the distribution of observations with the existence of unobserved subpopulations. Suppose for subjects $i = 1, \ldots, n$, the corresponding random vectors of observed data are $\mathbf{X}_1, \ldots, \mathbf{X}_n$, with realizations $\mathbf{x}_1, \ldots, \mathbf{x}_n$. We further suppose that there are C subpopulations in the overall population, and that each subpopulation is associated with a different distribution for the observed data, $f_c(\mathbf{X})$. Then, under the basic form of the mixture model, the probability density of subject i's observed responses $\mathbf{X}_i$ is

$$f(\mathbf{X}_i = \mathbf{x}_i) = \sum_{c=1}^{C} \pi_c * f_c(\mathbf{X}_i = \mathbf{x}_i), \tag{3.3}$$

where π_c is the probability that subject i is in subpopulation c, also known as the mixing weight, and $\sum_c \pi_c = 1$. A comprehensive overview of finite mixture models can be found in McLachlan and Peel (2004).

Mixture models have been widely used in psychometrics research. In fact, one of the most commonly used item response models, the three-parameter logistic model (Birnbaum, 1968),

$$P(X = 1 \mid \theta) = c + (1 - c) * \frac{1}{1 + \exp(-a(\theta - b))}, \tag{3.4}$$

can be reformulated into a mixture model,

$$P(X = 1 \mid \theta) = \frac{1}{1 + \exp(-a(\theta - b))} + c * (1 - \frac{1}{1 + \exp(-a(\theta - b))}), \tag{3.5}$$

where the probability of correct response is 1 with probability $\frac{1}{1+\exp(-a(\theta-b))}$, and the probability of correct response is c (guessing) with probability $1-\frac{1}{1+\exp(-a(\theta-b))}$. Mixture models have also been applied in educational measurement to address some practical issues in testing, such as identifying rapid-guessing or aberrant behaviors among test-takers (e.g., C. Wang & Xu, 2015), to identify compromised test items (e.g., McLeod, Lewis, & Thissen, 2003), and to model test speededness in time-constrained testing scenarios (e.g., Bolt, Cohen, & Wollack, 2002).

From this general formulation of mixture models, we can quickly see that the restricted latent class model in Chapter 1 and the hidden Markov learning models discussed in Chapter 2 can be deemed as mixture models. Under the RLCM, the probabilities of responding correctly on the assessment items depends on the attribute pattern of the subject. The unobserved latent attribute profiles can be regarded as the subpopulations, associated with mixing weights $\boldsymbol{\pi}$. Under the hidden Markov models for learning, the initial population membership of the subjects affect their transition probabilities to other attribute patterns at subsequent time points as well as their responses (and response times if applicable), and thus, the initial attribute patterns could also be regarded as subpopulations. However, we note that for the mixture hidden Markov models to be introduced next, the mixture refers to the subpopulations of the overall HMM instead of just the initial classifications.

Mixture HMMs. Langeheine and Van de Pol (1990) and Van de Pol and Langeheine (1990) proposed the mixed Markov latent class model, which, in its most general form, is the mixture of several first order hidden Markov models. Specifically, we index the subpopulations by $c = 1, \ldots, C$, the possible states of the Markov chain by $s \in \{1, \ldots, S\} = \mathcal{A}$, and the time points by $t = 1, \ldots, T$. Let the observed responses at time t by $\mathbf{X}_t$, then under the mixed Markov latent class model, the probability of the observed response sequence $\mathbf{X}_1, \ldots, \mathbf{X}_T$ is given by

$$P(\mathbf{X}_1, \ldots, \mathbf{X}_T) = \sum_{c=1}^{C} \pi_c \sum_{\forall (s_1,\ldots,s_T) \in \mathcal{A}^T} \left[\delta_{s_1|c} P(\mathbf{X}_1 \mid s_1, c) \prod_{t=2}^{T} \tau_{s_t|s_{t-1},c} P(\mathbf{X}_t \mid s_t, c) \right], \qquad (3.6)$$

where π_c is the probability of being in subpopulation c, $(s_1, \ldots, s_T) \in \mathcal{A}^T$ is any latent trajectory from time 1 to time T in the state space $\mathcal{A}$, $\delta_{s_1|c}$ is the initial probability of being in state s_1 for someone in subpopulation c, $\tau_{s_t|s_{t-1,c}}$ is the probability of transitioning from state s_{t-1} to state s_t for someone in subgroup c, and lastly, $P(X_t \mid s_t, c)$ is the probability of responding X_t given latent state s_t, for someone in subpopulation c. Model (3.6) assumes that different subpopulations different in all three components of the hidden Markov model, namely initial probabilities (δs), transition probabilities (τs), and response distributions ($P(\mathbf{X}_t \mid \cdot)$).

Vermunt, Tran, and Magidson (2008) further extended the mixed Markov latent class model in Equation (3.6) to incorporate time-invariant or time-dependent covariates, denoted $\mathbf{z}_{i,t}$ for subject i at time t. Denoting the subpopulation membership of subject i by W_i, which takes values in $1, \ldots, C$, the model can be obtained with a slight modification to Equation (3.6),

$$P(\mathbf{X}_1, \ldots, \mathbf{X}_T \quad | \quad \mathbf{z}_i) = \sum_{c=1}^{C} P(W_i = c \mid \mathbf{z}_i) \tag{3.7}$$
$$\times \sum_{\forall (s_1,\ldots,s_T) \in \mathcal{A}^T} \left[\delta_{s_1|c,\mathbf{z}_{i,1}} P(\mathbf{X}_1 \mid s_1, c, \mathbf{z}_{i,1}) \prod_{t=2}^{T} P(s_t \mid s_{t-1}, c, \mathbf{z}_{it}) P(\mathbf{X}_t \mid s_t, c, \mathbf{z}_{it}) \right].$$

This model replaces the constant subgroup probabilities, initial population membership probabilities, transition probabilities, and response probabilities in the mixed Markov latent class model by more general distributions, with the covariates $\mathbf{z}_i$.

3.2 Mixture Learning Model with Response Times and Response Accuracy

The mixture HMM formulation in (3.7) allows the initial hidden state distributions, transition model, and response model to differ across people in different subpopulations. It is

possible that the measurement or transition models among subgroups come from the same distribution but have group-specific parameters, and it is also possible that these models take different functional forms across different groups. We adopt a similar framework for modelling the learners' behaviors in a learning process. However, instead of assuming sub-populations of learners, we instead assume that at each time point, a learner can be in different modes. Below, we provide an example model, under which at each time point, a learner make take two possible learning modes, namely engaged ($D_{i,t} = 0$) or disengaged ($D_{i,t} = 1$).

For the mixture learning model with engaged and disengaged learning modes, we assume that under different learning modes, learners employ different solution strategies, and their probabilities of transitioning from non-mastery to mastery are different. Specifically, we assume that a learner in the engaged learning mode engages in solution behavior on the assessment items, by employing relevant skills to respond to the questions as accurately and quickly as possible. In this case, the learner's responses and response times can be modelled using the DINA response model in Equation (1.2) and the log-normal response times model in Equation (3.1), respectively. Similar to S. Wang et al. (2018), we consider three different versions of $G_{i,j,t}$, namely

- $G_{i,j,t} = t_{i,j}/T$;
- $G_{i,j,t} = \begin{cases} 1, & \text{if } \boldsymbol{\alpha}_{i,t} \succeq \mathbf{q}_j, \\ 0, & \text{otherwise.} \end{cases}$; and
- $G_{i,j,t} = \log(\sum_{m<t}(1 - D_{i,m}) \sum_h \eta^*_{i,h,m} + \sum_{q<j} \eta^*_{i,q,t} + 1)$, where $\eta^*_{i,h,m} = (\prod_{k=1}^K \alpha_{i,m,k}^{q_{h,k}}) \cdot \mathcal{I}(\mathbf{q}_h^T \mathbf{q}_j > 0)$.

The first two versions of $G_{i,j,t}$ are the same as in S. Wang et al. (2018), for which the interpretations are provided in the previous section. Note that for the third version of $G_{i,j,t}$, a slight modification is made in the calculation of the amount of practice. Instead of including all of the items that the subject has answered previously as a master, only the

ones s/he answered as a master in the engaged learning mode are included. As an example, we consider a learner who was disengaged at time 2 and answered all questions at that time point with rapid-guessing. Suppose at time 3, he switches back to the engaged learning mode and answers each item with a solution strategy. The items he answered at time 2, which he rapid-guessed, will not contribute to the amount of practice he has done when we model his response times at time 3.

In terms of the transition probability, we make the assumption that a learner in the engaged mode also has high engagement level in the learning process and thus may improve in skill mastery at that time point. In the engaged learning mode, the learner's transitions of attribute pattern at that time point is hence modelled using the higher-order hidden Markov CDM (HO-HM CDM) in Equation (2.1). Similar to the modeling for $G_{i,j,t}$, we make a slight modification to the practice term in the HO-HM CDM:

$$\text{logit}[P(\alpha_{i,k,t+1} = 1 \mid \alpha_{i,k,t} = 0, \boldsymbol{\alpha}_{i,t})] = \lambda_0 + \theta_i + \lambda_1 \sum_{\forall k' \neq k} \alpha_{i,k',t} + \lambda_2 \sum_{m=1}^{t} (1 - D_{i,m}) \sum_{j=1}^{J_t} q_{j,m,k}, \tag{3.8}$$

where the amount of practice, $\sum_{m=1}^{t}(1 - D_{i,m}) \sum_{j=1}^{J_t} q_{j,m,k}$, only includes previous items completed under the engaged mode. In addition, we note that the slope in front of θ_i is dropped in Equation (3.8). In this case, θ_i will be treated as a random intercept, whose variance is freely estimated, and its distribution is modelled together with τ_i to capture the population-level relationship between latent speed and latent learning ability.

On the other hand, if at a specific time point, a learner is in the disengaged learning mode, we assume this learner takes rapid-guessing strategies on assessment items and shows low engagement in the learning process. We model their rapid guessing strategy using similar methods as in C. Wang and Xu (2015), where the probability of correctly responding to item j is equal to some $g^* \in (0, 1)$ across all items, and the distribution of response times under

the rapid guessing strategy is also assumed to be the same across items, specifically,

$$\log(L_{i,j,t}) \mid D_{i,t} = 1 \sim \mathrm{N}(\mu_1, \sigma_1^2), \tag{3.9}$$

where μ_1 and σ_1^2 are the mean and variance of the log-response times in the disengaged mode. Lastly, the disengagement in the learning process is reflected in the transition probabilities from the current stage to the next. In other words, if a learner i is in the disengaged mode at time t, his or her attribute pattern at time $t+1$ is assumed to the unchanged from $\boldsymbol{\alpha}_{i,t}$. As a summary, Table 3.2 presents the learning, response, and response time models for the learners under the two different learning modes.

Table 3.1: Components of the mixture learning model with disengaged and engaged test takers.

Learning Mode	**engaged** ($D_{i,t} = 0$)	**Disengaged** ($D_{i,t} = 1$)
$P(\boldsymbol{\alpha}_{i,t+1} \mid \boldsymbol{\alpha}_{i,t}) =$	$\text{logit}[P(\alpha_{i,k,t+1} = 1 \mid \alpha_{i,k,t} = 0, \boldsymbol{\alpha}_{i,t})] =$ $\lambda_0 + \lambda_1\theta_i + \lambda_2 \sum_{\forall k' \neq k} \alpha_{i,k',t} +$ $\lambda_3 \sum_{m=1}^{t}(1 - D_{i,m}) \sum_{j=1}^{J_t} q_{j,m,k}$	$\mathcal{I}(\boldsymbol{\alpha}_{i,t+1} = \boldsymbol{\alpha}_{i,t})$
$P(X_{i,j,t} = 1) =$	$(1 - s_j)^{\prod_{k=1}^{K} \alpha_{i,t,k}^{q_{j,k}}} g_j^{1-\prod_{k=1}^{K} \alpha_{i,t,k}^{q_{j,k}}}$	g^*
$\log(L_{i,j,t}) \sim$	$N\left(\gamma_j - (\tau_i + \phi_0 * G_{i,j,t}), \frac{1}{a_j^2}\right)$	$N(\mu_1, \sigma_1^2)$

3.2.1 Bayesian Model Formulation

We set up our mixture learning model with engaged and disengaged modes under a Bayesian framework. Let $D_{i,t}$ denote the membership of subject i at time t in terms of whether s/he is disengaged, where $D_{i,t} = 1$ if i is disengaged at time t, and $D_{i,t} = 0$ otherwise. We assume that

$$D_{i,t} \sim \text{Bernoulli}(\omega), \tag{3.10}$$

where ω is the probability an arbitrary learner belongs to the disengaged group, and the prior distribution of ω is

$$\omega \sim \text{Beta}(1, 1). \tag{3.11}$$

The initial attribute pattern of learner i is assumed to be a multinomial sample from all $C = 2^K$ possible classes, with

$$P(\boldsymbol{\alpha}_{i,1} = \alpha_c) = \prod_{c=1}^{C} \pi_c^{\mathcal{I}(\boldsymbol{\alpha}_{i,1}=\boldsymbol{\alpha}_c)}, \tag{3.12}$$

where a Dirichlet prior distribution for the initial probabilities of each attribute pattern is used,

$$\boldsymbol{\pi} = [\pi_1, \ldots, \pi_C] \sim \text{Dirichlet}(1, \ldots, 1). \tag{3.13}$$

At time $t \in \{1, \ldots, T-1\}$, if a learner is in the engaged learning mode with $D_{i,t} = 0$, his or her attribute pattern at the next time point, $\boldsymbol{\alpha}_{i,t+1}$, conditioning on the attribute pattern at time t is modelled using the higher-order hidden Markov CDM in equation (2.1). Similar to S. Wang et al. (2018), we used the following prior probabilities for the learning model parameters:

$$\lambda_0 \sim \text{Normal}(0, 1), \lambda_2 \sim \text{Lognormal}(-1, 0.6), \lambda_3 \sim \text{Lognormal}(-1, 0.6). \tag{3.14}$$

And if the learner is disengaged at time t with $D_{i,t} = 1$, $\boldsymbol{\alpha}_{i,t+1} \mid D_{i,t} = 1$ is equal to $\boldsymbol{\alpha}_{i,t}$ with probability 1.

The responses of a learner under the engaged mode is assumed to follow the DINA model in Equation (1.2). A Beta prior was used for the slipping and guessing parameters of all the items, in other words,

$$p(s_j, g_j) \propto s_j^{a_s-1}(1-s_j)^{b_s-1} g_j^{a_g-1}(1-g_j)^{b_g-1} \mathcal{I}(0 \leq g_j < 1 - s_j \leq 1). \tag{3.15}$$

On the other hand, the response to an item j by a learner in the disengaged mode is assumed to be a Bernoulli sample with success probability g^*, in other words, $P(X_{i,j,t} = 1 \mid D_{i,t} = 1) = g^*$, where g^* is assumed to have a Beta$(1, 1)$ prior distribution.

At each time point $t = 1, \ldots, T$, if $D_{i,t} = 0$, subject i's response times on each item

follows the log-normal distribution in Equation (3.1). Similar as in S. Wang et al. (2018), we use the following priors for the response time model parameters:

$$\gamma_j \sim N(0,1), \phi_0 \sim N(0,1), \text{ and } a_j^2 \sim \text{Gamma}(1,1). \tag{3.16}$$

If $D_{i,t} = 1$, the reaction times to each item by learner i are assumed to follow the log-normal distribution given in Equation (3.9), with the following priors for the response time model parameters:

$$\mu_1 \sim N(0,1), \text{ and } \sigma_1^2 \sim \text{Inv-Gamma}(1,1). \tag{3.17}$$

Lastly, for each learner, his or her latent learning ability θ_i and initial latent speed τ_i in the engaged mode are assumed to follow a multivariate normal distribution, i.e., $(\theta_i, \tau_i)' \sim \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma})$, where the prior of the covariance matrix $\boldsymbol{\Sigma}$ follows the Inverse-Wishart distribution, in other words,

$$\boldsymbol{\Sigma} \sim \text{Inv-Wishart}(3, I_2). \tag{3.18}$$

3.2.2 Bayesian Full Conditional Distribution

Under the Bayesian modelling framework described above, the full likelihood of the subjects' responses and response times, as well as their speed, learning ability, and attribute patterns and learning modes at each time point conditioning on the fixed model parameters is given

by

$$
\begin{aligned}
& P(\mathbf{X}, \mathbf{L}, \boldsymbol{\alpha}, \boldsymbol{\theta}, \boldsymbol{\tau}, \mathbf{D} \mid \boldsymbol{\lambda}, \mathbf{s}, \mathbf{g}, g^*, \mathbf{a}, \boldsymbol{\gamma}, \phi_0, \mu_1, \sigma_1^2, \omega, \boldsymbol{\pi}, \boldsymbol{\Sigma}) \\
= & \prod_{i=1}^{N} \Bigg\{ p(\theta_i, \tau_i \mid \boldsymbol{\Sigma}) p(\boldsymbol{\alpha}_{i,t} \mid \boldsymbol{\pi}) \times \\
& \prod_{t=1}^{T-1} \Big[P(D_{i,t} \mid \omega) P(\mathbf{X}_{i,t}, \mathbf{L}_{i,t} \mid D_{i,1}, \ldots, D_{i,t}, \cdot) P(\boldsymbol{\alpha}_{t+1} \mid D_{i,1}, \ldots, D_{i,t}, \cdot) \Big] \times \\
& P(D_{i,T} \mid \omega) P(\mathbf{X}_{i,T}, \mathbf{L}_{i,T} \mid D_{i,1}, \ldots, D_{i,T}, \cdot) \Bigg\},
\end{aligned} \tag{3.19}
$$

where

$$
\begin{aligned}
& P(\mathbf{X}_{i,t}, \mathbf{L}_{i,t} \mid D_{i,1}, \ldots, D_{i,t}, \cdot) \\
= & \begin{cases}
\prod_{j=1}^{J_t} g^{*X_{i,j,t}} (1 - g^*)^{1 - X_{i,j,t}} f(L_{i,j,t} \mid \mu_1, \sigma_1^2), & \text{if } D_{i,t} = 1 \\
\prod_{j=1}^{J_t} (1 - s_j)^{\prod_{k=1}^{K} \alpha_{i,t,k}^{q_{j,k}}} g_j^{1 - \prod_{k=1}^{K} \alpha_{i,t,k}^{q_{j,k}}} f(L_{i,j,t} \mid \gamma_j, \tau_i, \phi_0, \boldsymbol{\alpha}_i, a_j), & \text{if } D_{i,t} = 0,
\end{cases}
\end{aligned} \tag{3.20}
$$

and

$$
P(\boldsymbol{\alpha}_{i,t+1} \mid D_{i,1}, \ldots, D_{i,t}, \cdot) = \begin{cases}
\mathcal{I}(\boldsymbol{\alpha}_{i,t+1} = \boldsymbol{\alpha}_{i,t}), & \text{if } D_{i,t} = 1, \\
\prod_{k=1}^{K} P(\alpha_{i,t+1,k} \mid \boldsymbol{\alpha}_{i,t}, \boldsymbol{\lambda}, \theta_i), & \text{if } D_{i,t} = 0.
\end{cases} \tag{3.21}
$$

Note that when $D_{i,t} = 0$, the transition probability and the response time distribution depend on $D_{i,1}, \ldots, D_{i,t-1}$ through the practice terms. When $D_{i,t} = 1$, neither the transition probability nor the response times depend on previous practice, so $P(\mathbf{X}_{i,t}, \mathbf{L}_{i,t} \mid D_{i,1}, \ldots, D_{i,t}, \cdot)$ and $P(\boldsymbol{\alpha}_{i,t+1} \mid D_{i,1}, \ldots, D_{i,t}, \cdot)$ reduce to $P(\mathbf{X}_{i,t}, \mathbf{L}_{i,t} \mid D_{i,t}, \cdot)$ and $P(\boldsymbol{\alpha}_{i,t+1} \mid D_{i,t}, \cdot)$.

We next present the conditional distribution of each model parameter given the other parameters and the observed responses and response times, which can be used to obtain random samples from the posterior distribution with a Gibbs sampling algorithm.

At each time point t and for each subject i, the conditional distribution of $D_{i,t}$ is

$$P(D_{i,t} = 1 \mid \omega, \mathbf{X}_{i,t}, \mathbf{L}_{i,t}, \boldsymbol{\alpha}_i) = \frac{\tilde{\pi}_{i,t,1}}{\sum_{d=0}^{1} \tilde{\pi}_{i,t,d}}. \tag{3.22}$$

When $G_{i,j,t} = t_{i,j}/T$ or $G_{i,j,t} = \mathcal{I}(\boldsymbol{\alpha}_{i,t} \succeq \mathbf{q}_j)$,

$$\tilde{\pi}_{i,t,d} = \begin{cases} P(D_{i,t} = d \mid \omega) \prod_{t^*=t}^{T-1} P(\boldsymbol{\alpha}_{i,t^*+1} \mid D_{i,1:(t-1)}, D_{i,t} = d, D_{i,t+1:t^*}\cdot) & \\ \times P(\mathbf{X}_{i,t}, \mathbf{L}_{i,t} \mid D_{i,t} = d, \cdot), & \text{if } t < T, \\ P(D_{i,t} = d \mid \omega) P(\mathbf{X}_{i,t}, \mathbf{L}_{i,t} \mid D_{i,t} = d, \cdot), & \text{if } t = T. \end{cases} \tag{3.23}$$

When $G_{i,j,t} = \log(\sum_{m<t}(1 - D_{i,m}) \sum_h \eta^*_{i,h,m} + \sum_{q<j} \eta^*_{i,q,t} + 1)$, $P(\mathbf{X}_{i,t}, \mathbf{L}_{i,t} \mid D_{i,t} = d, \cdot)$ in both cases of Equation (3.23) is replaced with $\prod_{t^*=t}^{T} P(\mathbf{X}_{i,t^*}, \mathbf{L}_{i,t^*} \mid D_{i,1:t-1}, D_{i,t} = d, D_{i,t+1:t^*}\cdot)$ to account for the effect of $D_{i,t}$ on the "practice effect" in the response times at later time points.

Here, $P(\mathbf{X}_{i,t^*}, \mathbf{L}_{i,t^*} \mid D_{i,1:t-1}, D_{i,t} = d, D_{i,t+1:t^*}, \cdot)$ is the probability of observing responses and latency $\mathbf{X}_{i,t^*}$ and $\mathbf{L}_{i,t^*}$ at time t^* given the D_is up to time t^*, with $D_{i,t}$ equal to d. And similarly, $P(\boldsymbol{\alpha}_{i,t^*+1} \mid D_{i,1:t-1}, D_{i,t} = d, D_{i,t+1:t^*}\cdot)$ is the probability that subject i takes attribute pattern $\boldsymbol{\alpha}_{i,t^*+1}$ at time $t^* + 1$, given D_is up to time t^* with $D_{i,t} = d$. These could be obtained based on Equations (3.20) and (3.21).

The conditional distribution of the mixture weight, ω, is

$$\omega \mid \mathbf{D} \sim \text{Beta}(1 + \sum_{i=1}^{N}\sum_{t=1}^{T} D_{i,t}, 1 + \sum_{i=1}^{N}\sum_{t=1}^{T}(1 - D_{i,t})). \tag{3.24}$$

For each subject i and each time point t, the conditional probability that the current attribute pattern $\boldsymbol{\alpha}_{i,t}$ equals $\boldsymbol{\alpha}_c \in \{0, 1\}^K$ is

$$P(\boldsymbol{\alpha}_{i,t} = \boldsymbol{\alpha}_c) = \frac{\tilde{\pi}_{ict}}{\sum_{c'=1}^{2^K} \tilde{\pi}_{ic't}}. \tag{3.25}$$

When $G_{i,j,t} = t_{i,j}/T$,

$$\tilde{\pi}_{ict} = \begin{cases} \pi_c P(\boldsymbol{\alpha}_{i,t+1} \mid \boldsymbol{\alpha}_{i,t}, D_{i,t}) P(\mathbf{X}_{i,t} \mid \boldsymbol{\alpha}_{i,t}, D_{i,t}), & \text{if } t = 1, \\ P(\boldsymbol{\alpha}_{i,t} \mid \boldsymbol{\alpha}_{i,t-1}, D_{i,t-1}) P(\boldsymbol{\alpha}_{i,t+1} \mid \boldsymbol{\alpha}_{i,t}, D_{i,t}) P(\mathbf{X}_{i,t} \mid \boldsymbol{\alpha}_{i,t}, D_{i,t}), & \text{if } 1 < t < T, \\ P(\boldsymbol{\alpha}_{i,t+1} \mid \boldsymbol{\alpha}_{i,t}, D_{i,t}) P(\mathbf{X}_{i,t} \mid \boldsymbol{\alpha}_{i,t}, D_{i,t}), & \text{if } t = T. \end{cases} \tag{3.26}$$

When $G_{i,j,t} = \mathcal{I}(\boldsymbol{\alpha}_{i,t} \succeq \mathbf{q}_j)$,

$$\tilde{\pi}_{ict} = \begin{cases} \pi_c P(\boldsymbol{\alpha}_{i,t+1} \mid \boldsymbol{\alpha}_{i,t}, D_{i,t}) P(\mathbf{X}_{i,t} \mid \boldsymbol{\alpha}_{i,t}, D_{i,t}) f(\mathbf{L}_{i,t} \mid \boldsymbol{\alpha}_{i,t}, D_{i,t}), & \text{if } t = 1, \\ P(\boldsymbol{\alpha}_{i,t} \mid \boldsymbol{\alpha}_{i,t-1}, D_{i,t-1}) P(\boldsymbol{\alpha}_{i,t+1} \mid \boldsymbol{\alpha}_{i,t}, D_{i,t}) P(\mathbf{X}_{i,t} \mid \boldsymbol{\alpha}_{i,t}, D_{i,t}) & \\ \times f(\mathbf{L}_{i,t} \mid \boldsymbol{\alpha}_{i,t}, D_{i,t}), & \text{if } 1 < t < T, \\ P(\boldsymbol{\alpha}_{i,t} \mid \boldsymbol{\alpha}_{i,t-1}, D_{i,t-1}) P(\mathbf{X}_{i,t} \mid \boldsymbol{\alpha}_{i,t}, D_{i,t}) f(\mathbf{L}_{i,t} \mid \boldsymbol{\alpha}_{i,t}, D_{i,t}), & \text{if } t = T. \end{cases} \tag{3.27}$$

And when $G_{i,j,t} = \log(\sum_{m<t}(1-D_{i,m}) \sum_h \eta^*_{i,h,m} + \sum_{q<j} \eta^*_{i,q,t} + 1)$, we replace $f(\mathbf{L}_{i,t} \mid \boldsymbol{\alpha}_{i,t}, D_{i,t})$ in Equation (3.27) with $\prod_{t^*=t}^{T} P(\mathbf{L}_{i,t^*} \mid \boldsymbol{\alpha}_{i,1:t-1}, \boldsymbol{\alpha}_{i,t} = \boldsymbol{\alpha}_c, \boldsymbol{\alpha}_{i,t+1:t^*}, D_{i,1:t^*}, \cdot)$ to account for the effect of previous attribute trajectories on the practice term in response times at later time points.

For the population proportions of the attribute patterns at time 1, $\boldsymbol{\pi}$, the conditional distribution is

$$\boldsymbol{\pi} \mid \boldsymbol{\alpha}_{1,1} \ldots, \boldsymbol{\alpha}_{N,1} \sim \text{Dirichlet}(1 + \tilde{N}), \tag{3.28}$$

where $\tilde{N} = [\sum_{i=1}^{N} \mathcal{I}(\boldsymbol{\alpha}_{i,1} = \boldsymbol{\alpha}_1), \ldots, \sum_{i=1}^{N} \mathcal{I}(\boldsymbol{\alpha}_{i,1} = \boldsymbol{\alpha}_{2^K})]$.

For any learner i, the conditional distribution of (θ_i, τ_i) is

$$P(\theta_i, \tau_i \mid \boldsymbol{\Sigma}, \boldsymbol{\alpha}_i, \mathbf{L}_i) \propto p(\theta_i, \tau_i \mid \boldsymbol{\Sigma}) \Big[\prod_{\substack{t<T: \\ D_{i,t}=0}} P(\boldsymbol{\alpha}_{i,t+1} \mid \boldsymbol{\alpha}_{i,t}, \theta_i, \boldsymbol{\lambda}) \Big] \Big[\prod_{t: D_{i,t}=0} f(\mathbf{L}_{i,t} \mid \tau_i, \boldsymbol{\gamma}, \mathbf{a}, \phi_0, \boldsymbol{\alpha}_i) \Big]. \tag{3.29}$$

And the conditional distribution of the covariance matrice of (θ_i, τ_i), $\boldsymbol{\Sigma}$, is

$$\boldsymbol{\Sigma} \mid \boldsymbol{\theta}, \boldsymbol{\tau} \sim \text{Inv-Wishart}\Big(N + 3, I_2 + \sum_{i=1}^{N} \begin{pmatrix} \theta_i^2 & \theta_i\tau_i \\ \theta_i\tau_i & \tau_i^2 \end{pmatrix}\Big). \tag{3.30}$$

For the slopes and intercept of the HO-HM CDM, $\boldsymbol{\lambda}$, the conditional distribution is

$$p(\boldsymbol{\lambda}) \prod_{i=1}^{N} \prod_{\substack{t<T-1: \\ D_{i,t}=0}} P(\boldsymbol{\alpha}_{i,t+1} \mid \boldsymbol{\alpha}_{i,t}, \boldsymbol{\lambda}, \theta_i). \tag{3.31}$$

The conditional distribution of the DINA model s_j, g_j for each item j is given by

$$P(s_j, g_j \mid \mathbf{X}_j, \boldsymbol{\alpha}, \mathbf{D}) \propto s_j^{\tilde{a}_s - 1}(1 - s_j)^{\tilde{b}_s - 1} g_j^{\tilde{a}_g - 1}(1 - g_j)^{\tilde{b}_g - 1} \mathcal{I}(g_j < 1 - s_j), \tag{3.32}$$

with

$$\tilde{a}_s = 1 + \sum_{\substack{i: D_{i,t}=0 \\ \& X_{i,j}=0}} \eta_{i,j,t}, \qquad \tilde{b}_s = 1 + \sum_{\substack{i: D_{i,t}=0 \\ \& X_{i,j}=1}} \eta_{i,j,t},$$

$$\tilde{a}_g = 1 + \sum_{\substack{i: D_{i,t}=0 \\ \& X_{i,j}=1}} (1 - \eta_{i,j,t}), \qquad \tilde{b}_g = 1 + \sum_{\substack{i: D_{i,t}=0 \\ \& X_{i,j}=0}} (1 - \eta_{i,j,t}),$$

where $\eta_{i,j,t}$ denotes the ideal response under the DINA model.

The conditional distribution of correct response probability for learners in the disengaged mode, g^*, is

$$g^* \mid \mathbf{X}, \mathbf{D} \sim \text{Beta}(\tilde{a}_{g^*}, \tilde{b}_{g^*}), \tag{3.33}$$

where

$$a_{g^*} = 1 + \sum_{i,t: D_{i,t}=1} \sum_{j=1}^{J_t} X_{i,j,t}, \qquad b_{g^*} = 1 + \sum_{i,t: D_{i,t}=1} \sum_{j=1}^{J_t} (1 - X_{i,j,t}).$$

For each item j, the conditional distribution of the time discrimination parameter a_j^2 is

$$a_j^2 \sim \text{Gamma}\Big(1 + \frac{\sum_{i=1}^{N}(1 - D_{i,t_{ij}})}{2}, 1 + \frac{\sum_{i=1}^{N}(1 - D_{i,t_{ij}})(\log L_{i,j,t_{ij}} + \tau_i + \phi_0 G_{i,j,t_{ij}} - \gamma_j)^2}{2}\Big), \tag{3.34}$$

where t_{ij} denotes the time at which item j is given to subject i. And the conditional distribution of the time intensity parameter, γ_j, is

$$\gamma_j \mid \mathbf{L}_j, \mathbf{D}, a_j, \phi_0, \boldsymbol{\tau} \sim N(\tilde{\mu}_\gamma, \tilde{\sigma}_\gamma^2), \text{ with} \tag{3.35}$$

$$\tilde{\sigma}_\gamma^2 = 1/\Big(1 + a_j^2 \sum_{i=1}^{N}(1 - D_{i,t_{ij}})\Big),$$
$$\tilde{\mu}_\gamma = \tilde{\sigma}_\gamma^2 * \Big\{a_j^2 \sum_{i=1}^{N}(1 - D_{i,t_{ij}})(\log L_{i,j,t_{ij}} + \tau_i + \phi_0 G_{i,j,t_{ij}})\Big\}.$$

For ϕ_0, the slope for the covariate describing speed increase over time in the engaged learning mode, the conditional distribution is

$$\phi_0 \mid \mathbf{L}, \boldsymbol{\alpha}, \boldsymbol{\tau}, \mathbf{a}, \boldsymbol{\gamma}, \mathbf{D} \sim N(\tilde{\mu}_\phi, \tilde{\sigma}_\phi^2), \text{ with} \tag{3.36}$$

$$\tilde{\sigma}_\phi^2 = 1/\Big(1 + \sum_{i=1}^{N}\sum_{t=1}^{T}(1 - D_{i,t_{ij}}) \sum_{j=1}^{J_t} a_j^2 G_{i,j,t_{ij}}^2\Big),$$
$$\tilde{\mu}_\phi = \tilde{\sigma}_\phi^2 * \Big\{\sum_{i=1}^{N}\sum_{t=1}^{T}(1 - D_{i,t_{ij}}) \sum_{j=1}^{J_t}\Big[a_j^2(\gamma_j - \tau_i - \log L_{i,j,t_{ij}})G_{i,j,t_{ij}}\Big]\Big\}.$$

Lastly, the conditional distributions of the mean and standard deviation of log-response times under the disengaged learning mode are as the following:

$$\mu_1 \mid \mathbf{L}, \mathbf{D}, \sigma_1^2 \sim N(\tilde{\mu}_{\mu_1}, \tilde{\sigma}_{\mu_1}^2), \text{ with} \tag{3.37}$$

$$\tilde{\sigma}^2_{\mu_1} = 1/\Big(1 + \frac{1}{\sigma^2_1} J_t \sum_{i=1}^{N}\sum_{t=1}^{T} D_{i,t}\Big),$$
$$\tilde{\mu}_{\mu_1} = \tilde{\sigma}^2_{\mu_1} * \Big\{\frac{1}{\sigma^2_1}\sum_{i=1}^{N}\sum_{t=1}^{T}\Big[D_{i,t}\sum_{j=1}^{J_t}\log L_{i,j,t}\Big]\Big\}. \text{ And}$$

$$\sigma^2_1 \mid \mathbf{L}, \mathbf{D}, \mu_1 \sim \text{Inv-Gamma}\Big(1+\frac{J_t \cdot \sum_{i=1}^{N}\sum_{t=1}^{T} D_{i,t}}{2}, 1+\frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\Big[D_{i,t}\sum_{j=1}^{J_t}(\log L_{i,j,t} - \mu_1)^2\Big]}{2}\Big). \tag{3.38}$$

A Gibbs sampler is developed to iteratively update the parameters above by sampling from their conditional distributions. For θ_i and for $\boldsymbol{\lambda}$, their conditional distributions do not resemble any known families of distributions, and thus, Metropolis-Hastings steps are used to update these parameters. A detailed discription of the MH sampling algorithm for these parameters can be found in S. Wang et al. (2018). Another thing to note is that when $D_{i,t,} = 1$, in other words, when a learner is disengaged, our model assumes that the attribute pattern at the next time point, $\boldsymbol{\alpha}_{i,t+1}$, is the same as $\boldsymbol{\alpha}_{i,t}$. In this case, $\boldsymbol{\alpha}_{i,t}$ and $\boldsymbol{\alpha}_{i,t+1}$ share the same unique attribute pattern. When we update the $\boldsymbol{\alpha}_{i,t}$s sequentially for each learner, instead of sampling each $\boldsymbol{\alpha}_{i,t}$ separately, sets of consecutive $\boldsymbol{\alpha}_i$s with no transitions in between (e.g., $\boldsymbol{\alpha}_{i,t}$ and $\boldsymbol{\alpha}_{i,t+1}$ if $D_{i,t} = 1$) are sampled together, conditioning on the previous attribute pattern (if available), the next unique attribute pattern (if available), and the observed responses and response times across all the time points sampled together (if available).

3.3 Simulation Study

We conducted a simulation study to evaluate the recovery of the mixture learning model parameters using the proposed Gibbs sampling algorithm. Out of the three versions of $G_{i,j,t}$ we discussed above, we used $G_{i,j,t} = \mathcal{I}(\boldsymbol{\alpha}_{i,t} \succeq \mathbf{q}_j)$ in the simulations as an illustration. However, the results should be generalizable to other types of Gs, because the only difference among the models using different types of Gs is the actual value of the covariate.

3.3.1 True Parameters

We simulated the attribute trajectories of $N = 500$ learners on $K = 4$ skills across $T = 5$ time points. The learners' initial attribute patterns were randomly sampled from the set of all possible attribute profiles ($\{0,1\}^K$) uniformly. And for each learner, their latent learning ability θ_i and latent speed τ_i were randomly generated from a multivariate normal distribution with mean $\mathbf{0}$ and covariance $\boldsymbol{\Sigma} = \begin{pmatrix} 3.24 & .36 \\ .36 & .25 \end{pmatrix}$.

At each time point, the learners were first administered $J_t = 10$ assessment items measuring their mastery on the 4 skills. And except for the last time point ($t = T$), after each block of assessment items, learners were administered a learning intervention, after which their attribute mastery status may change. To ensure a balanced sample for the estimation of the item parameters, an incomplete block design similar as in S. Wang et al. (2016) was used to counterbalance the order of item assignment to different learners.

At each time point $t = 1, \ldots, T$, the learners were randomly assigned to one of two possible learning modes, namely the engaged learning mode ($D_{i,t} = 0$) and the disengaged learning mode ($D_{i,t} = 0$). The true probability of $D_{i,t} = 1$ was set to be $\omega = .1$, in other words, at any time point, the probability that a learner is disengaged was 10%. Then, conditioning on the learner's mode at time t, the attribute mastery changes, responses, and response times were simulated with from different distributions. More specifically:

(1) **Transition.** If at time t, learner i is in the engaged learning mode ($D_{i,t} = 0$), the probability that the learner transitions from non-mastery to mastery on a skill is given by the modified HO-HM CDM in Equation (3.8). Similar to S. Wang et al. (2016), we assumed the monotonicity in the growth of attribute mastery, in other words, a mastered skill will not be forgotten. The true intercept (λ_0) and slopes (λ_1, λ_2) of the learning model were set to $-1, .3$, and $.05$, respectively. If learner i is disengaged at time t with $D_{i,t} = 1$, the learner's attribute pattern at the next time point, $\boldsymbol{\alpha}_{i,t+1}$, was set to be the same as the current one, $\boldsymbol{\alpha}_{i,t}$.

(2) **Response.** When a learner is in the engaged learning mode at time t ($D_{i,t} = 0$), the learner is assumed to engage in the solution behavior, and the responses were simulated under the DINA model in Equation (1.2). For each item in the testing item pool, the DINA model slipping parameters (s_j) were generated from Beta$(1, 10)$, and the guessing parameters (g_j) were generated from Beta$(2, 2)$. These Beta distribution parameters were selected to imitate the distribution of the estimated slipping and guessing probabilities from the Spatial Rotation Learning Program data set. On the other hand, if the learner is disengaged at time t with $D_{i,t} = 1$, a rapid guessing strategy is assumed and the learner's responses are generated from Bernoulli(.2), in other words, the probability of guessing correctly on any item was $g^* = .2$.

(3) **Response Times.** We assumed that when a learner is in the engaged learning mode, the observed response times follow the log-normal model in Equation (3.1), with $G_{i,j,t} = \mathcal{I}(\boldsymbol{\alpha}_{i,t} \succeq \mathbf{q}_j)$, which takes the value 1 if learner i has mastered all requisite skills for item j by time t and 0 otherwise. For each assessment item, the time intensity parameter γ_j was generated from $N(4, .5)$, and the time discrimination parameter a_j was generated from $U(2, 4)$. These distributions were selected to resemble the observed response time distributions in seconds from the Spatial Rotation Learning Program data if we assume a mean of 0 for the subjects' latent speeds. And for the slope in front of the covariate $G_{i,j,t}$, a true value of $\phi_0 = .3$ was used. In other words, we assumed a .3 increase in latent speed for any learner who has mastered all required skills of an item. If $D_{i,t} = 1$, in other words, learner i is disengaged at time t, the observed reaction times to any item at that time point was simulated from log-normal$(\mu_1 = 2, \sigma_1 = .5)$, translating to an average reaction time of approximately 8 seconds.

3.3.2 Parameter Estimation

To start the MCMC, we first generated initial values of all the model parameters, and each of them were sequentially updated given the others from the conditional distributions in the

section above. Specifically, the initial fixed parameters were generated as follows:

$$
\begin{array}{lll}
\lambda_0 \sim N(0,1), & \lambda_1 \sim U(0,1), & \lambda_2 \sim U(0,1), \\
\boldsymbol{\Sigma} = I_2, & \boldsymbol{\pi} \sim \text{Dirichlet}(\mathbf{1}), & \phi_0 \sim U(0,1), \\
\omega \sim U(0,.2), & g^* \sim U(0,.5), & s_j, g_j \sim U(0,.3), \\
\mu_1 \sim N(2,1), & \sigma_1 \sim U(0,1), & \gamma_j \sim N(3.45,.5^2), \\
a_j \sim U(2,4). & &
\end{array}
$$

The random parameters, namely $\mathbf{D}, \boldsymbol{\alpha}, \boldsymbol{\theta}$ and $\boldsymbol{\tau}$, were then randomly generated based on the corresponding fixed parameters.

A chain length of 15000 iterations was used for the MCMC, with the first 5000 as the burn-in that were excluded for the computation of the point estimates of the parameters. From the post burn-in iterations, we calculated the expected a posteriori (EAP) estimates of each of the model parameters by taking the average of the parameter samples. For the binary parameters, $\boldsymbol{\alpha}$ and $\mathbf{D}$, the final point estimates were dichotomized depending on whether the associated post burn-in average was less than or greater than .5.

3.3.3 Evaluation Criteria

The performance of the proposed algorithm is evaluated in terms of two aspects:

(1) MCMC chain convergence: To evaluate the model convergence, five separate chains with different starting values were run with chain lengths of 15000 iterations under each condition, based on one randomly simulated data set. The Gelman-Rubin proportional scale reduction factor (PSRF), commonly known as $\hat{R}$ (Gelman et al., 2014), was calculated for each parameter at different chain lengths from 5000 to 15000, with the first 5000 iterations as the burn-in, and the progression of the maximum $\hat{R}$ out of all estimated parameters is used to determine an adequate chain length for the MCMC algorithm.

(2) Parameter recovery: The ability of the proposed algorithm to accurately recover the true parameters was evaluated in terms of the following aspects. The recovery of attribute patterns of the students at each time point was evaluated using the attribute-wise agreement rate, $AAR = \frac{\sum_{i=1}^{N}\sum_{k=1}^{K}\mathcal{I}(\alpha_{ikt}=\hat{\alpha}_{ikt})}{N\times K}$, and the pattern-wise agreement rate, $PAR = \frac{\sum_{i=1}^{N}\mathcal{I}(\boldsymbol{\alpha}_{i,t}=\hat{\boldsymbol{\alpha}}_{i,t})}{N}$, between the true ($\boldsymbol{\alpha}$) and estimated ($\hat{\boldsymbol{\alpha}}$) attribute patterns. We further evaluated the recovery of $\phi_0, \boldsymbol{\Sigma}, \boldsymbol{\pi}$, $\boldsymbol{\lambda}$, ω, μ_1,σ_1, and g^* by comparing the mean and standard deviation of the posterior parameter samples to the true values. The recovery of the response model item parameters, $\mathbf{s}$ and $\mathbf{g}$, were evaluated in terms of their correlations with the true values and bias (Bias($\mathbf{s}$) $= \frac{\sum_{j=1}^{J_t\times T}(\hat{s}_j - s_j)}{Jt\times T}$, similarly for the $\mathbf{g}$s). The agreement between true and estimated response time model parameters ($\mathbf{a}$ and $\boldsymbol{\gamma}$), learning ability ($\boldsymbol{\theta}$) and latent speed ($\boldsymbol{\tau}$) were evaluated in terms of the correlation between true and estimated values, as well as the root mean squared error ($RMSE(\boldsymbol{\tau}) = \frac{\sum_{i=1}^{N}(\tau_i-\hat{\tau}_i)^2}{N}$, similarly for $\mathbf{a}, \boldsymbol{\gamma}, \boldsymbol{\theta}, \mathbf{s}$,and $\mathbf{g}$). Note that for each student, the data used to update θ are their transitions from non-mastery to either non-mastery or mastery. Therefore, once a student becomes a master of all skills, the subsequent $\boldsymbol{\alpha}$s will not be used to update θ, and no data on the transitions are available for students who have mastered all skills at the very beginning. For this reason, when computing the correlation between true and estimated learning abilities, we excluded the students who's estimated initial attribute pattern was $(1, 1, 1, 1)$. Finally, we computed the percentage of agreement between each true $D_{i,t}$ and estimated $\hat{D}_{i,t}$ to evaluate the sensitivity of the proposed algorithm in detecting whether a learner is engaged or disengaged at a given time point.

3.3.4 Results

Parameter Convergence. Figure 3.1 presents the change of the maximum univariate $\hat{R}$ among all model parameters as chain length increases. From the figure, we observe that after approximately 6000 iterations, the maximum $\hat{R}$ reduced to below 1.2, indicating parameter

convergence.

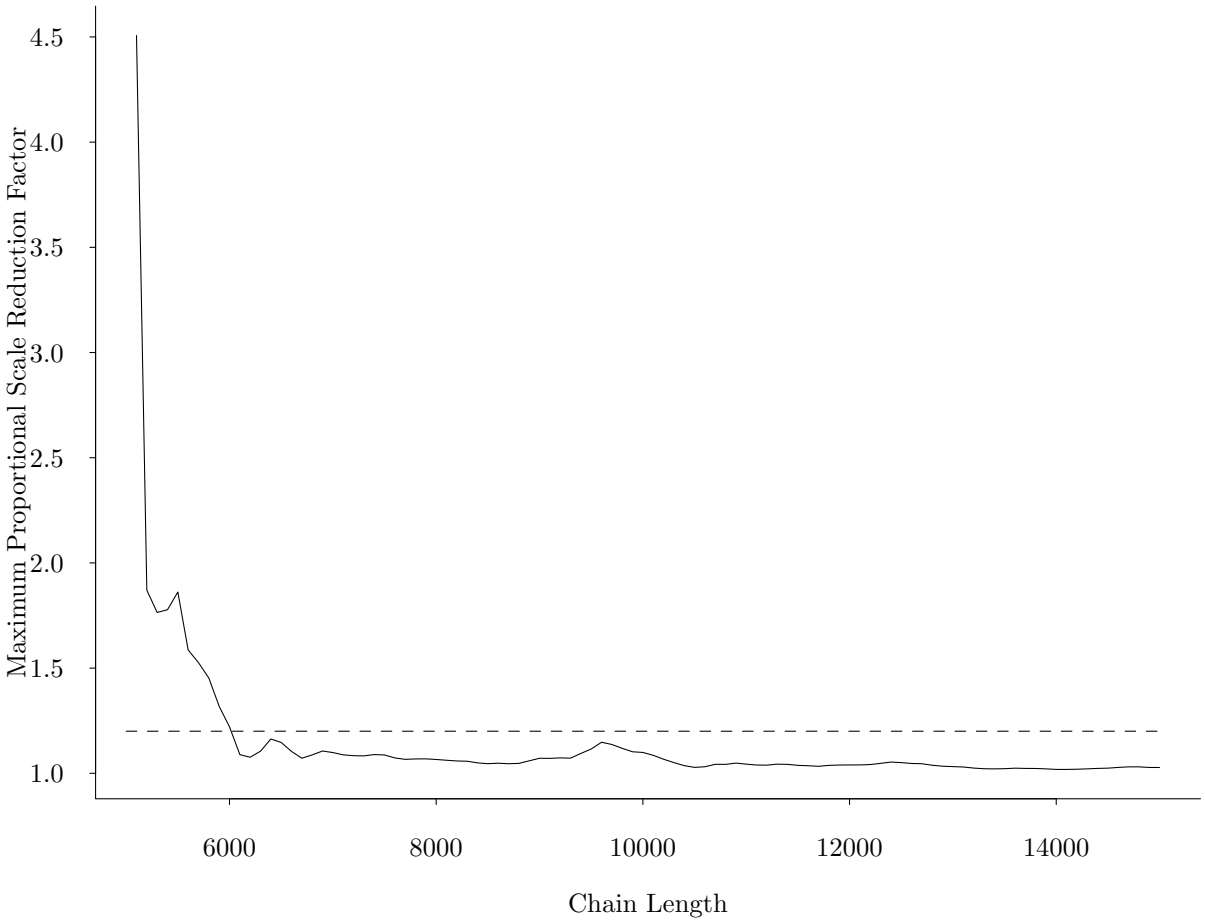

Figure 3.1: Maximum Brook-Gelman Proportional Scale Reduction Factor across all parameters with different chain lengths. The $x-$axis is the length of the MCMC chain, and the $y-$axis is the maximum PRSF. Dashed line represents the commonly used threshold of $\hat{R} = 1.2$ for parameter convergence.

Parameter Recovery. Table 3.2 presents the attribute-wise agreement rates (AARs) and the pattern-wise agreement rates (PARs) between the true and estimated attribute patterns ($\boldsymbol{\alpha}$) at each time point. Across all time points, the proposed estimation algorithm achieved over 88% accuracy in measuring the presence/absence of attributes for each participant. The estimation accuracy was the lowest for the initial time point ($t = 1$), and it increased as t increased, achieving over 96% agreement at $t = 5$.

Table 3.3 presents the true values, posterior means, and the posterior standard devi-

Table 3.2: The attribute-wise and pattern-wise agreement rates (AARs and PARs) between the true and estimated $\boldsymbol{\alpha}$.

Time	$t=1$	$t=2$	$t=3$	$t=4$	$t=5$
AAR	0.885	0.942	0.958	0.968	0.968
PAR	0.630	0.800	0.854	0.882	0.894

ations of the fixed parameters in the model, namely the variance of the general learning ability (σ_θ^2), covariance between learning ability and latent speed ($\sigma_{\theta\tau}$), variance of latent speed (σ_τ^2), the transition model's intercept (λ_0) and slopes (λ_1, λ_2), the probability of disengagement (ω), the coefficient for the increase of latent speed (ϕ_0) for engaged learners, the correct response probability in the disengaged mode (g^*), and the mean (μ_1) and standard deviation (σ_1) of the log response times in the disengaged mode. Most of the EAPs for these parameters were fairly close to the true values used for data generation, with small standard deviations. However, we observe that the parameters associated with the transition model ($\sigma_\theta^2, \boldsymbol{\lambda}$) showed larger biases and larger standard deviations. One possible reason is that with $T = 5$, each learner could be observed on at most 4 transitions, and considering that some learners started with mastery of all or most of the skills at the initial time point and that some learners might be disengaged at a selection of time points, the actual number of observations for transitions are usually less than 4 per learner. Thus, the amount of data available for estimating the transition model parameters, as well as the θs and their variance, is very limited.

The correlation between true and estimated values, as well as the RMSEs of $\mathbf{a}, \boldsymbol{\gamma}, \mathbf{s}, \mathbf{g}, \boldsymbol{\tau}$, and $\boldsymbol{\theta}$ are presented in Table 3.4. For the items' response time model parameters ($\mathbf{a}, \boldsymbol{\gamma}$), the DINA model parameters ($\mathbf{s}, \mathbf{g}$), and the learners' initial latent speeds ($\boldsymbol{\tau}$), there was a high agreement between the true and estimated values, with correlations over 97% and RMSEs less than .02. For the latent learning abilities of the learners ($\boldsymbol{\theta}$), the estimate values demonstrated larger errors with correlation around 75% and RMSE over 1.35. Similar to the larger errors in the transition model parameter estimates, we think the larger error in the estimation of $\boldsymbol{\theta}$ can potentially be attributed to the paucity of data available to update

Table 3.3: The true and estimated covariance between $\boldsymbol{\theta}$ and $\boldsymbol{\tau}$ ($\boldsymbol{\Sigma}$), learning model parameters ($\boldsymbol{\lambda}$), mixing weight (ω), coefficient for speed growth (ϕ_0) in the engaged learning mode, correct response probability of learners in the disengaged mode (g^*), and the log-normal mean (μ_1) and standard deviation (σ_1) of the response time distribution in the disengaged mode. "True" stands for the true value of the parameters, "EAP" is the average of the parameter samples across iterations, and "SD" is the standard deviation of the parameter samples across iterations.

	σ_θ^2	$\sigma_{\theta\tau}$	σ_τ^2	λ_0	λ_1	λ_2
True	3.24	0.360	0.250	-1.000	0.300	0.050
EAP	4.748	0.417	0.26	-0.596	0.210	0.106
SD	1.083	0.081	0.018	0.198	0.074	0.027
	ω	ϕ_0	g^*	μ_1	σ_1	
True	0.100	0.300	0.200	2.000	0.500	
EAP	0.092	0.292	0.205	2.005	0.493	
SD	0.006	0.006	0.008	0.010	0.007	

θ_i for each subject.

Table 3.4: Root mean square errors (RMSE) and correlations between true and estimated DINA item parameters ($\mathbf{s}, \mathbf{g}$), response time model parameters ($\mathbf{a}, \boldsymbol{\gamma}$), and latent speed ($\boldsymbol{\tau}$) and learning ability ($\boldsymbol{\theta}$) of learners.

Parameter	$\mathbf{a}$	$\boldsymbol{\gamma}$	$\mathbf{s}$	$\mathbf{g}$	$\boldsymbol{\tau}$	$\boldsymbol{\theta}$
Correlation	0.985	1.000	0.983	0.973	0.995	0.755
RMSE	0.015	0.016	$<.001$	0.002	0.018	1.351

Finally, across several repetitions of the simulation study, the estimated learning mode of each subject at each time point, $D_{i,t}$, showed high agreement with the true values, with the proportion of correctly estimated entries in $\mathbf{D}$ very close to ($> 99.9\%$), if not equal to, 100%. This suggests that under the proposed estimation algorithm, whether a learner is disengaged or engaged at a given time point could be detected correctly most of the times based on their response times, responses, and transitions in attribute mastery.

3.4 Discussion

In the current chapter, we proposed a mixture learning model with two possible learning modes, namely the engaged mode and the disengaged mode. Under different modes, learners

are assumed to demonstrate different learning and response behaviors, leading to differences in the distributions of attribute mastery transitions over time, item responses, and response times. A Bayesian Gibbs sampling algorithm was proposed to estimate the parameters of the mixture model, and simulation studies showed that the model parameters could be accurately estimated, the learners' learning mode could be detected with high accuracy, and the chains converged with as little as 6000 iterations.

When heterogeneity exists in the learning process, failure to account for different types of learning modes could lead to inaccurate estimates of many parameters. As a simple illustration we fitted the response and response times data generated in the simulation study to the joint learning model of responses and response times under the HO-HM CDM framework proposed by S. Wang et al. (2018), which assumes all learners are in the engaged mode across all time points, with response distribution, response times distribution, and transition model same as those under the engaged mode in the mixture model. Results suggested a remarkable decrease in estimation accuracy of the attribute patterns (drop of average AAR from .944 to .697), latent learning abilities (drop of correlation from .755 to $-.020$), latent speeds (drop of correlation from .995 to .649), DINA model item parameters ($\mathbf{s}$: correlation drop from .983 to .696, $\mathbf{g}$: from .973 to .737), and the items' time discrimination parameters (drop of correlation from .985 to .291). Thus, when the empirical data suggest the existence of disengaged learning for some learners at some occasions, fitting the proposed mixture learning model instead of a homogeneous learning model could possibly improve the accuracy in parameter estimates.

The proposed mixture learning model has the potential to detect student disengagement in an online learning context. Compared to traditional classroom learning, online learning programs often provide the students with a significantly more flexible and less controlled environment. Whereas teachers in traditional classrooms can directly observe the students' behaviors and their reactions to different interventions, in online learning, the educators do not interact face to face with the students. The proposed mixture learning model framework

provides a way for educators to infer the online learners' learning mode (e.g., engaged or disengaged) based on the observed responses and reaction times to assessment questions at different time points. And the learners' reaction times to assessment questions, in addition to the responses, provide an extra source of information in the estimation of not only the learners' attribute mastery, but also their learning modes at each learning stage. For instance, a disengaged learner and a engaged learner struggling with the contents may both have low accuracy on their responses to the assessment items, and by looking merely at a response vector, it is hard to infer if a learner is putting in efforts yet struggling or not paying attention at all. If we look at the reaction times to the assessment items together with the response accuracy, a fast yet incorrect response would be indicative of rapid guessing, which could be more likely for the disengaged learners compared to struggling engaged learners. If the educators can detect when a learner is showing low engagement, targeted stimulation can be provided to the disengaged learner to increase their engagement level, which will in turn increase their gain from the online learning experience.

The simulation study conducted here is just a first step at evaluating the performance of the mixture learning model and the estimation algorithm, and it has a lot of limitations. To name a few, more systematic simulation studies should be conducted with multiple repetitions, as well as different conditions of true parameter distribution and model specification, such as the type of $G_{i,j,t}$ used to explain reaction time decrease over time. In addition, in the current simulation study, only one mixing weight (ω) and one type of distribution of the responses and response times of the disengaged learners were considered. The assumed distribution of the reaction times of the disengaged learners is also relatively restrictive, with a small log-mean and a relatively small log-variance. By imposing this distribution of the response times, we essentially assume that disengaged learners employ the rapid guessing strategy on their item responses. However, it is possible that in practice, some of the disengaged learners may not be responding rapidly to assessment items, especially when other distractions are available. In the presence of larger heterogeneity in reaction times under

the disengaged learning mode, the performance of the proposed mixture model in detecting learner disengagement is worth evaluating. The robustness of the mixture model in the case of no mixtures should also be examined by evaluating the model parameter recovery and the estimate for $D_{i,t}$s when the true generating model does not assume the presence of disengagement in the learning process.

A lot of follow up research could be done along the lines of mixture learning models. As an immediate next step, the proposed model will be applied to the data collected from the Spatial Rotation Learning Program (S. Wang et al., 2016), where the raw reaction times of the learners suggested that a lot of learners with low response accuracy also responded the quickest among all participants. The fit of the proposed model, compared to a the joint learning model of responses and response times (S. Wang et al., 2018), will be evaluated. The proposed mixture model with engaged and disengaged learners also has a lot of room for extensions. For example, instead of modeling the learning mode as a Bernoulli random variable with a fixed probability of disengagement, a higher order model could be used to describe the probability that a learner is disengaged at a specific time point, given a set of time dependent or time independent covariates, such as learners' demographic information or other characteristics, the mode of instruction (e.g., video, text, interactive exercise), or the temporal position of the current learning block (e.g., first learning block which may show slow warm-up of the learners, or later learning blocks on which learners may demonstrate fatigue), et cetera. The mixture model with two learning modes can also be extended to include three or more possible learning modes, which could be used to differentiate different types of disengagement or to capture other learning behaviors other than engaged and disengaged, such as a warm-up mode, where students have low familiarity with the learning environment and need some time to adjust before fully engaging.

Chapter 4

`hmcdm`: An R Package for Fitting Learning Models

4.1 Introduction

With the increased prevalence of online learning systems and the recent advocates for integrating assessments with instructions (U.S. Department of Education, 2016), psychometrics researchers became increasingly interested in adapting the methods in measurement research to aid learning (e.g., Chang, 2015). With the ability to assess the mastery on fine-grained skills, cognitive diagnosis models (Rupp et al., 2010) can be readily applied to the longitudinal learning setting, to track students' acquisition of knowledge or skills over time, identify the covariates affecting learning outcome, and understand students' learning behaviors.

`hmcdm`, which stands for hidden Markov cognitive diagnosis modeling, is an R package for fitting longitudinal models for learning under the cognitive diagnosis framework, including the higher-order hidden Markov model (S. Wang et al., 2016), the first order hidden Markov model (Chen, Culpepper, Wang, & Douglas, 2017), the reduced-RUM and NIDA learning models discussed in Chapter 2, and the joint model for learning with responses and response times (S. Wang et al., 2018). The package allows users to simulate item responses (and response times if applicable) under several learning models, to fit the models using Markov Chain Monte Carlo (MCMC) methods, to compute point estimates of parameters based on the MCMC samples, and to evaluate and compare different models using Deviance Information Criterion (Celeux, Forbes, Robert, & Titterington, 2006) and posterior predictive probabilities (Sinharay et al., 2006). In addition to the functions for modeling learning, users can also access the data from the Spatial Rotation Learning Program (S. Wang et al.,

2016).

4.2 Availability

A free copy of the `hmcdm` package, including the C++ source codes, the Spatial Rotation data, a user manual, and an R script of vignettes demonstrating the use of the functions in the package are available to the public on GitHub. Users can access the package by downloading the binary source from `https://github.com/tmsalab/hmcdm/releases` and installing the package in RStudio from package archive files. Alternatively, users can also install the package by first installing the `devtools` package in R, and then typing `devtools::install_github('tmsalab/hmcdm')` in the R console. The package can be installed on Windows, Linux, and macOS operating systems.

4.3 Documentations

A PDF manual with complete documentations of all data objects, callable functions, as well as examples for each function can be found in the source of the package. Example R code for response simulation, model estimation, and model fit analyses using each learning model mentioned above are also available in the `hmcdm/R` folder in the package source.

References

Baker, R. S., Gowda, S. M., Wixon, M., Kalka, J., Wagner, A. Z., Salvi, A., ... Rossi, L. (2012). Towards sensor-free affect detection in cognitive tutor algebra. *International Educational Data Mining Society*.

Beck, J. E. (2004). Using response times to model student disengagement. In *Proceedings of the its2004 workshop on social and emotional intelligence in learning environments* (pp. 13–20).

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. *Statistical theories of mental test scores*.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture rasch model with ordinal constraints. *Journal of Educational Measurement*, *39*(4), 331–348.

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*(4), 434–455.

Celeux, G., Forbes, F., Robert, C. P., & Titterington, D. M. (2006). Deviance information criteria for missing data models. *Bayesian analysis*, *1*(4), 651–673.

Chang, H.-H. (2015). Psychometrics behind computerized adaptive testing. *Psychometrika*, *80*(1), 1–20.

Chen, Y., Culpepper, S., Shiyu, W., & Douglas, J. (2017). A hidden markov model for learning trajectories with application to spatial rotation skills. *Applied Psychological Measurement*. doi: 10.1177/0146621617721250

Chen, Y., Culpepper, S. A., Wang, S., & Douglas, J. (2017). A hidden markov model for learning trajectories in cognitive diagnosis with application to spatial rotation skills. *Applied Psychological Measurement*, 0146621617721250.

Chen, Y., Liu, J., Xu, G., & Ying, Z. (2015). Statistical analysis of Q-matrix based diagnostic classification models. *Journal of the American Statistical Association*, *110*(510), 850–866.

Chiu, C.-Y., Douglas, J. A., & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika*, *74*(4), 633–665.

Chiu, C.-Y., & Köhn, H.-F. (2015). The reduced RUM as a logit model: Parameterization and constraints. *Psychometrika*, 1–21.

Collins, L. M., & Wugalter, S. E. (1992). Latent class models for stage-sequential dynamic latent variables. *Multivariate Behavioral Research*, *27*(1), 131–157.

Culpepper, S. A. (2015). Bayesian estimation of the DINA model with Gibbs sampling. *Journal of Educational and Behavioral Statistics*, *40*(5), 454–476.

Culpepper, S. A., & Hudson, A. (2017). An improved strategy for bayesian estimation of the reduced reparameterized unified model. *Applied Psychological Measurement*, 0146621617707511.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*(2), 179–199.

Eddelbuettel, D., François, R., Allaire, J., Chambers, J., Bates, D., & Ushey, K. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, *40*(8), 1–18.

Felder, R. M., & Silverman, L. K. (1988). Learning and teaching styles in engineering education. *Engineering education*, *78*(7), 674–681.

Gabrielsen, A. (1978). Consistency and identifiability. *Journal of Econometrics*, *8*(2), 261–263.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (Vol. 2). CRC press Boca Raton, FL.

George, A. C., Robitzsch, A., Kiefer, T., Groß, J., & Ünlü, A. (2016). The R package CDM for cognitive diagnosis models. *Journal of Statistical Software*, *74*(2), 1–24. doi: 10.18637/jss.v074.i02

Hartz, S. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality* (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign.

Henrie, C. R., Halverson, L. R., & Graham, C. R. (2015). Measuring student engagement in technology-mediated learning: A review. *Computers & Education*, *90*, 36–53.

Henson, R. A., & Templin, J. (2007). Large-scale language assessment using cognitive diagnosis models. In *Annual Meeting of the National Council on Measurement in Education, Chicago, IL.*

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.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*(3), 258–272.

Langeheine, R. (1988). Manifest and latent markov chain models for categorical panel data. *Journal of Educational Statistics*, *13*(4), 299–312.

Langeheine, R., & Van de Pol, F. (1990). A unifying framework for markov modeling in discrete space and discrete time. *Sociological Methods & Research*, *18*(4), 416–441.

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. *Studies in Social Psychology in World War II Vol. IV: Measurement and Prediction*, 362–412.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on tatsuoka's rule-space approach. *Journal of Educational Measurement*, *41*(3), 205–237.

Li, F., Cohen, A., Bottge, B., & Templin, J. (2015). A latent transition analysis model for assessing change in cognitive skills. *Educational and Psychological Measurement*, *76*(2), 181–204.

McLachlan, G., & Peel, D. (2004). *Finite mixture models.* John Wiley & Sons.

McLeod, L., Lewis, C., & Thissen, D. (2003). A bayesian method for the detection of item preknowledge in computerized adaptive testing. *Applied Psychological Measurement*, *27*(2), 121–137.

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R News*, *6*(1), 7–11. Retrieved from `http://CRAN.R-project.org/doc/Rnews/`

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. Guilford Press.

Sinharay, S., & Almond, R. G. (2007). Assessing fit of cognitive diagnostic models: A case study. *Educational and Psychological Measurement*, *67*(2), 239–257.

Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement*, *30*(4), 298–321.

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

Templin, J. L., & Hoffman, L. (2013). Obtaining diagnostic classification model estimates using Mplus. *Educational Measurement: Issues and Practice*, *32*(2), 37–50.

U.S. Department of Education. (2016). *The 2016 National Educational Technology Plan.* Retrieved from `https://tech.ed.gov/netp/`.

Van de Pol, F., & Langeheine, R. (1990). Mixed markov latent class models. *Sociological methodology*, 213–247.

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, *31*(2), 181–204. (doi:10.3102/10769986031002181)

Vermunt, J. K., Tran, B., & Magidson, J. (2008). Latent class models in longitudinal research. *Handbook of longitudinal research: Design, measurement, and analysis*, 373–385.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, *61*(2), 287–307.

Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology*, *68*(3), 456–477.

Wang, S., Yang, Y., Culpepper, S. A., & Douglas, J. A. (2016). Tracking skill acquisition

with cognitive diagnosis models: A higher-order, hidden markov model with covariates. *Journal of Educational and Behavioral Statistics*, 1076998617719727.

Wang, S., Zhang, S., Douglas, J., & Culpepper, S. (2018). Using response times to assess learning progress: A joint model for responses and response times. *Measurement: Interdisciplinary Research and Perspectives*, *16*(1), 45–58.

Xu, G. (2017). Identifiability of restricted latent class models with binary responses. *The Annals of Statistics*, *45*(2), 675–707.

Xu, G., & Zhang, S. (2016). Identifiability of diagnostic classification models. *Psychometrika*, *81*(3), 625–649.

Yoon, S. Y. (2011). *Psychometric properties of the revised Purdue Spatial Visualization Tests: Visualization of rotations (the revised PSVT-R)* (Unpublished doctoral dissertation). Purdue University.

Zhang, S., & Chang, H.-H. (2016). From smart testing to smart learning: how testing technology can assist the new generation of education. *International Journal of Smart Technology and Learning*, *1*(1), 67–92.