A DIAGNOSTIC CLASSIFICATION MODEL FOR POLYTOMOUS ATTRIBUTES

by

YU BAO

(Under the Direction of Laine P. Bradshaw)

ABSTRACT

Diagnostic classification models (DCMs) are statistical models designed to provide feedback about students' understandings of multiple latent knowledge components, termed *attributes*. Compared to traditional measurement models that place students' abilities on a unidimensional continuous scale, DCMs classify students into levels of attribute mastery and can achieve high reliability with shorter assessments than those required by continuous measurement. To this point, however, DCMs have been used to provide dichotomous feedback about students' mastery and non-mastery levels. In educational contexts, further delineating mastery categories may be useful for meaningfully grouping students to provide tailored instruction or interventions.

To identify additional mastery levels, we extended the current DCM framework by developing a polytomous DCM (PDCM) that classifies students into more than two mastery levels for each attribute. In the PDCM, we defined a polytomous attribute as an ordinal latent variable, and we allowed the item response probabilities to vary differentially between different mastery levels. A constrained PDCM was proposed in this dissertation by constraining some item parameters to be equal to reduce the number of item parameters and required fewer items and smaller sample size compare to the PDCM.

Two simulation studies were conducted to investigate the model estimation and model misspecification. The first study examined the attribute classification accuracies and the item parameter estimation across various conditions. The results shown the PDCM required longer test lengths to yield accurate classification for the attributes and item parameter estimation. The second study evaluated model misspecification. When the attribute mastery levels were under-specified, examinees in the intermediate mastery groups were forced to be classified into other mastery groups and thus the feedback provided was less detailed. When the attribute mastery levels were over-specified, most examinees were still classified into original mastery level groups.

An empirical study was conducted to illustrate the application of the PDCM using data from an assessment designed for special education students which measured four mathematics problem solving skills. We compared PDCM results with a dichotomous DCM framework which shown the PDCM provided improved model-data fit. The more detailed attribute classification also illustrated the utility of PDCM feedback for education practitioners.

INDEX WORDS: diagnostic classification models, latent class analysis, polytomous attribute, model misspecification

A DIAGNOSTIC CLASSIFICATION MODEL FOR POLYTOMOUS ATTRIBUTES

By

YU BAO

B.S., Beijing Normal University, China, 2010

M.S., Beijing Normal University, China, 2013

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2019

A DIAGNOSTIC CLASSIFICATION MODEL FOR POLYTOMOUS ATTRIBUTES

By

YU BAO

Major Professor: Laine Bradshaw
Committee: Allan Cohen
Shiyu Wang
Scott Ardoin

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
May 2019

DEDICATION

To my parents, Li Li and Hanwu Bao.

ACKNOWLEDGEMENTS

I first would like to express my sincere gratitude to my academic advisor, Dr. Laine Bradshaw, who opens the world of psychometrics and DCMs for me. I am thankful to her patience, enthusiasm, and understanding which helped me learn everything from scratch. Without such a brilliant scholar, I would not complete this dissertation study successfully.

I would like to thank my committee members, Drs. Allan Cohen, Shiyu Wang, and Scott Ardoin for their supports in all aspects to this dissertation. I appreciate their time in meeting with me and providing valuable comments, which helped me improve this study and pushed me go further than I thought.

I also would like to thank all my friends and fellow graduate students for their valuable suggestions and company during the process. I would specially thank Drs. Mei Ling Ong and Man Kit Lee who offered kindness guidance in writing and warm meals when I have no time to cook.

I appreciate my parents, Li Li and Hanwu Bao, for always being around and ready to offer help. Thank you, my Mom, for flying from China and staying with me most of the time while I was writing the dissertation. Your healthy meals and encouragement had made the process much easier and more joyful.

TABLE OF CONTENTS

Page

APPENDICES

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

With the increasing need of using assessments to better support students' learning, many psychometric and statistical models have been used to assist test designing and score reporting. Particularly, more attention has been given to a group of psychometric models that provide multidimensional fine-grained diagnoses about students' knowledge or skills (Rupp, & Templin, 2008). The group of models are referred to as diagnostic classification models (DCMs; e.g., Rupp, Templin, & Henson, 2010), or the cognitive diagnostic models (CDMs; e.g., Leighton & Gierl, 2007). DCMs focus on classifying students into mastery or nonmastery levels of specific skills or knowledge components, which are usually termed as *attributes.* Teachers and parents can use this type of feedback to remedy the knowledge that students' need to improve in the future learning process. However, current studies on DCMs are limited to providing such dichotomous feedback. The dissertation focuses on generalizing dichotomous DCMs to polytomous DCMs to provide more detailed information from educational assessments. This chapter presents the motivation of the dissertation by introducing the utility of DCMs and the need for the development of the polytomous general DCM.

Recent legislation indicates the potential of applying DCMs to K-12 assessments. The "No Child Left behind Act" (2001; Section 1111, [b][3][c] xii) and the "Every Student Succeed Act" (2015; Section 1111, [b] [3] [c] x) both emphasize an assessment shall be used to

> produce individual student interpretive, descriptive, and diagnostic reports, consistent with clause (iii) that allow parents, teachers, and principals to understand and address the specific academic needs of students, and include information regarding achievement on academic assessments aligned with State academic achievement standards, and that are

> provided to parents, teachers, and principals, as soon as is practicably possible after the assessment is given, in an understandable and uniform format, and to the extent practicable, in a language that parents can understand

Moreover, the “Every Student Succeed Act” allows more flexibility to the states to administer assessments in that a single annual assessment can be divided into a set of smaller and more specific assessments (USA Today, Dec. 11, 2015; Darling-Hammond, Bae, Cook-Harvey, Lam, Mercer, Podolsky, & Stosich, 2016).

The recent education policies shed a light on the application of DCMs to the K-12 education to provide diagnostic reports for each individual student to meet the academic needs. The purpose of the dissertation study is to broaden the use of DCMs by proposing a general DCM framework that is appropriate to provide diagnostic feedback about not only the dichotomous mastery levels (mastery and nonmastery), but also the polytomous mastery levels for specific knowledge or skills. The model allows researchers and test administrators to obtain more flexible feedback on what attribute and to what extent a student needs to improve on the attribute. In the following sections, we compare traditional psychometric models and DCMs, as well as the idea of generalizing the dichotomous DCMs to the polytomous DCMs.

Traditional Psychometric Models

Classical test theory (CTT; e.g., Crocker & Algina, 1986) is a traditional test theory that assumes an examinee’s observed test score is the sum of the true test score and the error. The observed test score is on a continuous scale and represents an examinee’s ability, that is, a higher test score means a higher ability. When two or more examinees have the same observed test score, CTT treats the examinees as having equal ability without considering which items they answered incorrectly. However, an examinee who missed more easy items because of carelessness and answered more difficult items correctly might have a higher ability than other examinees who have the same total score. Item difficulty is the percentage of the correct response for the item and the item discrimination is the point biserial correlation between the item responses and the observed test scores. These item statistics are highly sample dependent and cannot be compared under two samples, meaning if a population has a higher average test score than the other, it is likely the item is easier, and the item difficulty is higher.

Rooted in the CTT, item response theory (IRT; e.g., Hambleton, Swaminathan, & Rogers, 1991; Baker and Kim, 2004) considers an examinee's ability as a latent variable and places examinees' abilities on a unidimensional continuous scale. Item response theory is a framework that contains many statistical models to estimate the probability of answering an item correctly given an examinee's ability. Different from CTT, the IRT models are not sample dependent meaning even if two examinee groups took the test in different time and locations, the examinees can be equated on the same continuous scale. However, the test conducted under the IRT framework requires items to measure a unidimensional latent ability. An item that measures more than one latent ability is often treated as a misfitting item and needs to be removed from the test. Though the unidimensional IRT framework can be generalized to the multidimensional IRT (MIRT; Reckase, 2009), a test under the MIRT framework requires much longer test length and large sample size to yield accurate estimate of examinees' multidimensional latent abilities (Jiang, Wang, and Weiss, 2016).

Diagnostic Classification Models

The development of the diagnostic assessments is based on the limitations of the current unidimensional testing theories. Different from CTT and IRT that rank students on a unidimensional continuous latent scale, DCMs focus on classifying students into mastery or nonmastery levels of knowledge components. These knowledge components may be specific content areas or curriculum standards. DCMs are useful for the formative assessments because DCMs provide feedback about multiple fine-grained knowledge without the requirement of long test length and large sample size. Recent studies have shown that DCMs performed uniformly greater reliability for examinees' latent attribute estimation compared to IRT under the same test length condition (Templin and Bradshaw, 2013).

In DCMs, the latent knowledge or skills are called attributes. If the mastery levels of attributes are dichotomous, the value of the attributes are usually defined as 0 or 1, where 0 represents nonmastery and 1 represents mastery. Suppose a test measures *A* attributes, there are 2^A possible attribute profiles, and an examinee will be classified into one of the attribute profiles based on his/her item response pattern of the test. For example, if the examinee answered most of the items which measured Attribute a correctly, he/she would more likely be classified as a master of this Attribute a. In DCMs, we use a Q-matrix to represent the relationship between items and attributes. Suppose the test has *I* items, the Q-matrix is an I by *A* matrix with dichotomous entries, where 0 indicates the item does not measure the attribute and 1 indicates the item measures the attribute.

Many DCMs have been proposed in the past two decades. Each DCM describes different attribute behaviors on an item. In the dissertation, we choose to use a general DCM framework called the loglinear cognitive diagnosis model (LDCM; Henson, Templin, & Wilse, 2009) as it can obtain most sub models of DCMs by constraining the item parameters. The commonly used sub models includes the deterministic inputs, noisy, ‘‘and’’ gate (DINA; Haertel, 1989; Junker & Sijtsma, 2001) model; deterministic inputs, noisy, ‘‘or’’ gate (DINO; Templin & Henson, 2006) model; additive CDM (A-CDM; de la Torre, 2011); the linear logistic model (LLM; Maris, 1999); and the reduced reparameterized unified model (R-RUM; DiBello, Roussos, & Stout, 2007; Hartz, 2002).

We generalize the LCDM to a polytomous DCM framework and compared the results of using the LCDM with the polytomous DCM as a baseline model. The LCDM is an ANOVA like model that includes an intercept, main effects, and interactions to predict the log-odds of the probability of a correct answer for an item. The intercept is the log-odds of the probability of a correct answer when an examinee is nonmastery for all required attributes of the item; the main effect of an attribute is the increase of the log-odds of the item response probability when an examinee is a master of the attribute; and the interaction is the effect on the item response probability when an examinee possesses multiple required attributes by the item. Another benefit of the LCDM is that it is an item-level measurement model, which allows attributes to behave differently across items. Besides the LCDM, there are other general DCMs, such as the general deterministic-input, noisy-and-gate model *(G*-DINA; de la Torre, 2011), the general diagnostic model (GDM; von Davier, 2005), and the Generalized Diagnostic Classification Models for Multiple Choice Option-Based Scoring (GDCM-MC; DiBello, Henson, and Stout, 2015).

General DCMs have been generalized to model more complex scenarios. For example, the scaling individuals and classifying misconceptions model (SICM; Bradshaw and Templin, 2013) is a nominal response model that aims at detecting examinees' misconceptions and estimating examinees' general abilities simultaneously. The nominal response LCDM (NR LCDM; Bradshaw, 2011) is a simplified version of the SICM which only focuses on detecting examinees' misconceptions through nominal responses. The multiple-choice DINA (MC-DINA; de la Torre, 2009) models examinees' nominal responses of a multiple-choice item to detect examinees' mastery of knowledge or skills. The restricted sequential G-DINA (RS-GDINA; Ma, de la Torre, 2016) and unrestricted sequential G-DINA model (US-GDINA; Ma, de la Torre, 2016) can model the ordered polytomous item responses. Minchen et al. (2017) proposed the continuous DINA (c-DINA) to model the continuous item responses. Moreover, DCMs were also applied to the longitudinal studies to provide feedback about examinees' learning progress (Madison, 2016; Wang et al., in press). In this dissertation, we extend the general DCM to model polytomous item responses.

A Generalization of the Polytomous DCMs

We generalize DCMs which measure dichotomous attribute mastery levels to polytomous mastery levels. Prior similar studies have proposed DCMs for polytomous attributes including the polytomous RUM (Templin, 2004), the ordered-category attribute coding (OCAC; Tarelitz, 2004; Chen and de la Torre, 2011) framework, and the general diagnostic model (GDM; von Davier, 2005). These models are described in detail in Chapter 3. This dissertation study proposes a general polytomous DCM, named PDCM, to guide DCM users to define attribute mastery levels, understand the measurement model, and interpret the results.

The PDCM defines attribute mastery levels as nonnegative integers from 0 to the highest mastery level, where 0 represents the nonmastery level and the largest value represents the highest mastery level. For example, if an attribute has four mastery levels, the values of the attribute is 0, 1, 2, and 3. The order of the values means the order of mastery levels but does not necessarily mean the difference between two mastery levels is equal to the difference between the two values. That is, the difference between nonmastery level to the first mastery level is not equal to 1, as well as not equal to the difference between the first mastery level to the second mastery level. The entries of the Q-matrix for the PDCM are dichotomous, where 0 represents the item does not measure the attribute and 1 represents the item measures the attribute.

The measurement model form of the PDCM is generalized from the LCDM, where the PDCM has the intercept, main effects, and the interactions. The difference between the PDCM and the LCDM in that the main effects and interactions are defined at the attribute mastery level. For example, if Item *i* measures attributes a_1 and a_2, and both attributes have three mastery levels, the PDCM for Item *i* has two main effects for Attribute a_1 and two main effects for Attribute a_2. The two main effects represent the increase of the log-odds of the item response probability when an examinee has mastered the attribute one level higher. The interactions represent the influence on the item response probability between the different combinations of the attributes and their mastery levels. Therefore, the PDCM creates the most flexibility on the item response probabilities across all attribute profiles.

Another flexibility of the PDCM lies in the application of the PDCM. The PDCM does not constrain all attributes to have the same mastery levels. For example, assume a test measure three polytomous attributes. Stakeholders might pursue different mastery levels for each attribute such that the feedback provided for each attribute is customized. The PDCM can fulfill this need and can also suggest test administrators the most appropriate the mastery levels for each attribute through statistical tests. When the attributes have two mastery levels, the PDCM is equal to the LCDM.

Overview of Chapters

Chapter 1 provides the introduction of the diagnostic assessments and the purpose of proposing the general polytomous DCM. Chapter 2 introduces a general dichotomous DCM framework and summarized the existing studies of the polytomous DCMs. Chapter 3 proposes a general polytomous DCM, named the PDCM, by explaining how the two key components – measurement model and structural model, as well as a constrained polytomous DCM, named the cPDCM. Chapter 4 demonstrates the designs of two simulation studies and an empirical study. The two simulation studies are to investigate the item parameter estimation and classification accuracy, and the model misspecification. The empirical study aims to present an application of the saturated PDCM and the constrained PDCM to a real educational assessment to provide a guidance for the potential users of the PDCM and the cPDCM. Chapter 5 summarizes the results of the two simulation studies and the empirical study mentioned in Chapter 4. Chapter 6 is a conclusion of the dissertation and a discussion about future study direction.

CHAPTER 2

THEORETICAL BACKGROUND

Chapter 2 presents an introduction of a general dichotomous DCMs and an overview of the existing literature about DCMs measuring polytomous attributes. This chapter illustrates the dichotomous and polytomous DCMs by four factors: definition of attribute mastery levels, Q-matrix, Q-matrix, measurement model and structural model. The purpose of the chapter is to review the existing literature and summarize the key factors of both dichotomous and polytomous DCMs. The chapter provides a theoretical foundation of the proposed general polytomous DCM in the following chapters.

The Diagnostic Classification Models

Attribute Profile

Suppose a test measures *A* attributes, examinee *e*'s attribute profile is denoted as a categorical latent vector $\boldsymbol{\alpha_e} = (\alpha_{e1}, \cdots, \alpha_{eA})'$ where each entry of the vector represents the attribute mastery level. In most DCMs, the attribute mastery levels are dichotomous, where $\alpha_{ea} = 0$ represents Examinee *e* is not a master of Attribute a and $\alpha_{ea} = 1$ represents Examinee *e* is a master of Attribute a. Considering all the combinations of the dichotomous attribute mastery levels, examinees were into 2^{A} attribute profiles based on their item responses.

Q-matrix

We assume a test contains *I* items and measures *A* attributes. Each item measures one or more attributes. In this dissertation, we use an *I by A* matrix called a Q-matrix (Tatsuoka; 1990) to represent the test blueprint and indicate which attributes are measured by an item. The rows in the Q-matrix represent the items and the columns represent the attributes. The entries of the Q-matrix are 0s and 1s, where 0 represents that an attribute is not hypothesized to be measured by an item and 1 represents that an attribute is hypothesized to be measured by an item. For example, $q_{ia} = 0$ means item *i* does not measure attribute a, and $q_{ia} = 1$ means item *i* measures attribute *a*. More specifically, a vector $\boldsymbol{q_i} = (\mathrm{q}_{\mathrm{i}1}, \cdots, q_{iA})'$ in the Q-matrix indicates which attributes are measured by item *i*.

Table 2.1 illustrates an example Q-matrix in Bradshaw et al. (2014). The test contained 27 effective items and was to measure teachers' understanding of the fraction arithmetic. More specifically, the test measured 4 attributes named: reference units (RU), partitioning and iterating (PI), appropriateness (APP), multiplicative comparison (MC). The Q-matrix, in this example, is a 27 by 4 matrix with 0s and 1s as entries. Item 1 measured only one attribute, RU and Item 14 measured two attributes, RU and MC. In total, the four attributes were measured by 14, 10, 5 and 5 items respectively. Among the 27 items, there were 20 items that only measured one attribute, which were also referred to as simple items in this dissertation, and the remaining 7 items measured two attributes.

The test blueprint or Q-matrix is established during the test design. Specifying the Q-matrix is essential to make diagnostic inferences about whether a student has mastered the attributes an item is measuring. For example, if an item measures a single Attribute a, a student with a correct response for this item is more likely a master of Attribute a, while another student with an incorrect response is more likely a nonmaster of Attribute *a*.

Item Response

Examinee e's item response vector for the test is denoted as $\boldsymbol{X_e} = (x_{e1}, x_{e2}, \cdots, x_{eI})'$, where I is the total number of items for a test. In this study, we only consider dichotomous item responses, where $x_{ei} = 1$ means Examinee *e* has answered Item *i* correctly and $x_{ei} = 0$ means Examinee *e* has answered Item *i* incorrectly.

Overview of the Probability Structure of DCMs

DCMs parameterize the relationship of an item response and the attribute profiles. DCMs assume local independency at the item level, meaning the item responses of an examinee are conditionally independent given the examinee's attribute profile. Given local independence, the likelihood function includes the simple product of conditional item response probabilities across examinees. The probability of examinee *e*'s item response vector can be expressed as:

$$
\begin{aligned}
\mathrm{P}(\mathbf{X_e} = \boldsymbol{x_e}) &= \sum_{c=1}^{2^A} P(\boldsymbol{X_e} = \boldsymbol{x_e}, \boldsymbol{\alpha_e} = \boldsymbol{\alpha_c}) \\
&= \sum_{c=1}^{2^A} \nu_c P(\boldsymbol{X_e} = \boldsymbol{x_e} | \boldsymbol{\alpha_e} = \boldsymbol{\alpha_c}) \\
&= \sum_{c=1}^{2^A} \nu_c \prod_{i=1}^{I} P(X_{ei} = 1 | \boldsymbol{\alpha_c})^{x_{ei}} \big(1 - P(X_{ei} = 1 | \boldsymbol{\alpha_c})\big)^{1-x_{ei}} \qquad (2.1)
\end{aligned}
$$

Where $\mathrm{P}(\mathbf{X_e} = \boldsymbol{x_e})$ is the probability that examinee *e* has the item response vector $\boldsymbol{x_e}$, ν_c is the probability of examinee *e* is classified into attribute profile group *c* where *c* could be any number from 1 to 2^A. In DCMs, the component ν_c can then further be modeled with the attribute relationships and called the *structural model*. Four structural models were introduced in the current literature including *the log-linear model* (e.g., Henson & Templin, 2005; Xu & von

Davier, 2008a), *the unstructured tetrachoric model* (e.g., Hartz, 2002), *the structured tetrachoric model* (e.g., de la Torre & Douglas, 2004; Templin et al., 2007; Templin & Henson, 2009), and *the unstructured structural model* (e.g., Rupp, Templin and Henson, 2010). Applying structural models in the DCM framework can reduce the complexity of the parameterization of the latent attribute space.

The form of $P(X_{ei} = 1|\boldsymbol{\alpha_c})$ is called the *measurement model*, and it can be modeled with the deterministic inputs, noisy, ‘‘and’’ gate (DINA; Haertel, 1989; Junker & Sijtsma, 2001) model, deterministic inputs, noisy, ‘‘or’’ gate (DINO; Templin & Henson, 2006) model, additive CDM (A-CDM; de la Torre, 2011), the linear logistic model (LLM; Maris, 1999), reduced reparameterized unified model (R-RUM; DiBello, Roussos, & Stout, 2007; Hartz, 2002). In the following sections, the structural model and measurement model are explained in detail.

The Measurement Model: The Log-linear Cognitive Diagnostic Model (LCDM)

In this study, we first introduce an existing saturated measurement model for the saturated dichotomous DCM, called the log-linear cognitive diagnostic model (LDCM; Henson, Templin, & Wilse, 2009). The LCDM is an ANOVA-like item-level generalized linear and latent mixed model that contains intercepts, main effects of attributes, and interactions effects for combinations of attributes. It models an examinee *e’s* item response probability for item *i*, $P(X_{ei} = 1|\boldsymbol{\alpha}_e)$, as a monotonic increasing function of the attributes measured by item *i*, that is, examinee *e* is more likely to answer item *i* correctly when he/she masters more required attributes. The form of the model is

$$logit\ P(X_{ei} = 1|\boldsymbol{\alpha}_e) = \log\frac{P(X_{ei} = 1|\boldsymbol{\alpha}_e)}{P(X_{ei} = 0|\boldsymbol{\alpha}_e)} = \lambda_{i,0} + \boldsymbol{\lambda}_i^T \boldsymbol{h}(\boldsymbol{\alpha_e}, \boldsymbol{q_i}), \qquad (2.2)$$

where $\lambda_{i,0}$ is the intercept for item *i; $\boldsymbol{\lambda}_i^T$* is a vector that contains all the main effects and possible interactions; $\boldsymbol{h}(\boldsymbol{\alpha_e}, \boldsymbol{q_i})$ is the vector that contains all the combinations of the attribute profile elements and q-vector elements. More specifically, the right-side of the equation can be expanded as

$$\lambda_{i,0} + \boldsymbol{\lambda_i^T} \boldsymbol{h}(\boldsymbol{\alpha_e}, \boldsymbol{q_i}) = \lambda_{\mathrm{i,0}} + \sum_{a=1}^{A} \lambda_{i,1,(a)} \alpha_{ea} q_{ia} +$$

$$\sum_{a=1}^{A-1} \sum_{a'=a+1}^{A} \lambda_{i,2,(a,a')} \alpha_{ea} \alpha_{ea'} q_{ia} q_{ia'} + \cdots \qquad (2.3)$$

where $\lambda_{i,1,(a)}$ represents the main effect for attribute a, $\lambda_{i,2,(a,a')}$ represents the two-way interactions for attribute *a* and *a'*, and the ellipsis represents the higher-order interactions.

An intercept $\lambda_{\mathrm{i,0}}$ is the log-odds of the item response probability of a correct response for item *i* when examinee *e* is a nonmaster for all the attributes measured by item *i.* $\lambda_{\mathrm{i,0}}$ conceptually corresponds to the guessing effect for examinees who are nonmasters of the required attributes. $\lambda_{\mathrm{i,0}}$ can be any real number ranging from the negative infinity to the positive infinity. The negative infinity corresponds to the item response probability of 0, meaning examinees who are nonmasters have 0 probability of answering item *i* correctly, while positive infinity corresponds to the item response probability of 1, meaning examinees who are nonmasters will certainly answer item *i* correctly. A smaller $\lambda_{\mathrm{i,0}}$ represents a lower probability of answering item *i* correctly and a larger $\lambda_{\mathrm{i,0}}$ represents a higher probability of answering item *i* correctly. Especially, when $\lambda_{\mathrm{i,0}}$ equals 0, the item response probability for item *i* is .50. In a real testing scenario, an item with a large intercept is usually easier and causes the effects of attributes measured by the item to have a restricted range.

The main effect $\lambda_{i,1,(a)}$ is the increase of the log-odds of the item response probability when examinee *e* is a master of attribute *a* measured by item *i.* The product of the elements of attribute profile α_{ea} and q-vector q_{ia} indicates when the main effect $\lambda_{i,1,(a)}$ appears in the equation, that is, when $\alpha_{ea} = 1$ *and* $q_{ia} = 1$. Because the item response function is monotonically increasing on the mastery levels of the required attributes, the correct item response logit is constrained to be larger for an examinee mastering attribute a, $\lambda_{\mathrm{i},0} + \lambda_{i,1,(a)}$, than for an examinee who has not mastered attribute a, $\lambda_{\mathrm{i},0}$. Therefore, the main effects are constrained to be greater than zero, i.e., $\lambda_{i,1,(a)} > 0$. Given the intercept $\lambda_{\mathrm{i},0}$ for item *i*, a larger main effect represents a stronger effect of the required attribute, meaning a larger increase on the item response probability. In the real testing scenario, an item with larger main effects are preferred because it contributes more to classifying examinees into attribute mastery groups.

The two-way and the higher-order interactions allow more flexibilities on the item response probability under different attribute behaviors. The LCDM is a saturated model because all possible main effects and interactions are modeled. For example, if item *i* measures two attributes: Attribute 1 and 2, the LCDM for item *i* is

$$\log\frac{P(X_{ei} = 1|\boldsymbol{\alpha}_e)}{P(X_{ei} = 0|\boldsymbol{\alpha}_e)} = \lambda_{i,0} + \lambda_{i,1,(1)}\alpha_{e1} + \lambda_{i,1,(2)}\alpha_{e2} + \lambda_{i,2,(1,2)}\alpha_{e1}\alpha_{e2} \tag{2.4}$$

The two-way interaction term between Attribute 1 and 2, $\lambda_{i,2,(1,2)}$, appears in the equation when $\alpha_{e1}\alpha_{e2} = 1$, meaning examinee *e* is a master of both Attribute *a* and *a'* ($\alpha_{e1} = 1, \alpha_{e2} = 1$). If an item measures two attributes, there is only one two-way interaction in the model. By including the two-way interaction, the item response probabilities is more flexibly defined and the model allows more general increase of the log-odds of the item response probability for mastering both attributes compared to that for mastering one attribute. Because of the monotonicity assumption, we have

$$\lambda_{i,0} + \lambda_{i,1,(1)} + \lambda_{i,1,(2)} + \lambda_{i,2,(1,2)} > \lambda_{i,0} + \lambda_{i,1,(1)} \tag{2.5}$$

$$\lambda_{i,0} + \lambda_{i,1,(1)} + \lambda_{i,1,(2)} + \lambda_{i,2,(1,2)} > \lambda_{i,0} + \lambda_{i,1,(2)}$$

After simplification, the constraint for the two-way interaction is $\lambda_{i,2,(1,2)} > -\lambda_{i,1,(1)}$ and $\lambda_{i,2,(1,2)} > -\lambda_{i,1,(2)}$. Similarly, the higher-order interactions are the combinations of all the possible attributes measured by an item.

In summary, suppose an item measures *k* attributes, there are one intercept, *k* main effects, $\binom{k}{2}$ two-way interactions, $\binom{k}{3}$ three-way interactions, …, and $\binom{k}{k} = 1$ *k*-way interaction in the LCDM. Therefore, the total number of item parameters in the saturated LCDM equals $1 + \binom{k}{2} + \binom{k}{3} + \cdots + \binom{k}{k} = 2^k$.

Many DCMs have been proposed in the past decades. The reason we use the LCDM in this study is that most DCMs can be obtained from the LCDM by constraining some item parameters. For example, the DINA (Junker & Sijsma, 2001) model can be obtained by constraining the LCDM with the main effects and lower-order interactions equal to 0 and constraining the highest-order interaction to be positive. The DINO model can be obtained by constraining the item response probabilities for all attribute profiles except the nonmastery group to be equal and larger than the item response probability for the nonmastery group. The CRUM can be obtained by constraining all interaction terms to be equal to 0.

The Structural Model

Other than understanding the attribute-item relationship, researchers might also seek to know the correlations among attributes. Practically, the correlations among attributes are good indicators of the model dimensionality. For example, a correlation of around .70 between two attributes is considered to be a reasonable representation of multidimensionality of the attributes and a well-constructed latent test structure (Bradshaw et al., 2014). While a correlation is higher than .90, it might implicate the two attributes measure the same latent scale and we might need to combine the two attributes to reduce the dimensionality. Mathematically, modeling the correlations between attributes can improve the classification accuracy because the DCMs with only the measurement models are assumed to have independent attributes, which is not realistic in an educational assessment.

With the local independence assumption and the probability form of the measurement model, DCMs can classify examinees into one of the possible attribute profiles based on their item responses. For the overall population, we can obtain the proportion of the examinees being classified into each attribute profile group, respectively. Since the attribute mastery levels are dichotomous in the LCDM, we can treat the proportion of an attribute profile as the corresponding element of a $2 \times \cdots \times 2 = 2^{\mathrm{A}}$ table. For each pair of attributes, we can simply compute the marginal combinations of the mastery levels between these two attributes, which is a 2×2 table with mastery and nonmastery as columns for the first attribute and mastery and nonmastery as rows for the second attribute. We then can use the correlation coefficient for the dichotomous variables, called tetrachoric correlation, to compute the correlations between any pair of the attributes measured by the test.

Like the structural equation models, we need the *structural* components ν_c which is the probability of an examinee being classified into a specific attribute profile group to measure the correlations among attributes. For a test measures *A* attributes, there are $(2^A - 1)$ structural components with the constraint $\sum_{c=1}^{2^A} \nu_c = 1$ that need to be estimated in a DCM. To reduce the number of the strucrual parameters, ν_c can be further modeled under different probability forms, refered to *structural models* (Henson & Templin, 2005, 2006; Henson et al., 2009; Xu & von Davier, 2008a).

The structural model we use in this study is the log-linear model generalized from the categorical data analysis (Agresti, 2012) that treats the outcome variables as the group number and the predictors as the mastery levels of attributes corresponds to each group. The log-linear structural model contains the linear combination of the main effects and interactions for all the attributes being measured by the test as a kernel function with a log link function of the structural component $\mathrm{\nu_c}$.

$$\mu_c = \log \nu_c / \nu_{2^A} = \gamma_0 + \sum_{a=1}^{A} \gamma_{1,(a)} \alpha_{ca} + \sum_{a=1}^{A-1} \sum_{a'=a+1}^{A} \gamma_{2,(a,a')} \alpha_{ca} \alpha_{ca'} + \cdots \quad (2.6)$$

where $\mathrm{\mu_c}$ is the natural log of the ratio of $\mathrm{\nu_c}$ and $\mathrm{\nu_{2^A}}$ with the last attribute profile as the reference group. γ_0 is the intercept; $\gamma_{1,(a)}$ in the kernel function is the main effect for $\mathrm{\alpha_{ca}}$ which is the mastery level of Attribute a in the attribute profile *c*; $\gamma_{2,(a,a')}$ is the two-way interaction for α_{ca} and $\alpha_{ca'}$; the ellipsis means all the other higher-order interaction terms. Note that the value of the kernel function for the last attribute profile group is 0, that is $\mathrm{\mu_{2^A}} = \log \frac{\nu_{2^A}}{\nu_{2^A}} = \log 1 = 0$. Thus, we have the intercept equals the negative sum of all the main effects and interactions.

$$\gamma_{\mathrm{i},0} = -\sum_{a=1}^{A} \gamma_{1,(a)} - \sum_{a=1}^{A-1} \sum_{a'=a+1}^{A} \gamma_{2,(a,a')} - \cdots \quad (2.7)$$

Although in the different scale, the value of the kernel function μ_c represents the value of ν_c, where a larger μ_c means a higher probability of being classified into the attribute profile *c*. As mentioned before, because the sum of ν_c is 1, the transformation between μ_c and ν_c is

$$\nu_c = \frac{\exp(\mu_c)}{\sum_{c'=1}^{2^A} \exp(\mu_{c'})} \tag{2.8}$$

The saturated parameterization for the log-linear structural model has $(2^{\mathrm{A}} - 1)$ parameters including all the main effects and interactions, which is the same number of the structural components ν_c. Since the higher order interactions are hard to yield a significant difference from 0 under a limited sample size, we can reduce the number of parameters estimated by constraining some higher-order interactions equal to 0 with a reasonable loss of the model flexibility.

Existing DCMs for Polytomous Attributes

The Generalized Linear Mixed Proficiency Models

While most DCMs classify examinees into two latent classes which are usually nonmastery or mastery groups for an attribute, Templin (2004) pointed out the need of examinees being classified into three or more attribute mastery levels and proposed the Generalized Linear Mixed Proficiency Models (GLMPM) which generalized the latent attribute mastery levels from dichotomous to polytomous or even continuous estimate. The GLMPM contains the measurement model to relate observed responses to the multidimensional attributes and the structural model to characterize the correlations among the attributes or the continuous latent traits. The measurement model can be any model within the DCM framework to represent the probability of answering an item correctly when a student has a specific attribute profile. In the paper, the author used the Reparameterized United Model (RUM; Hartz, 2002) for dichotomous attribute levels and the generalized RUM to represent the probability of a correct response under the polytomous attribute. The response variable in the measurement model within the GLMPM framework remains dichotomous and the only change is the mastery levels of the attributes.

The attributes in the GLMPM are defined as ordered integers 0, 1, …. For example, if an attribute has three mastery levels, the values of the attribute are 0, 1 or 2. These values may represent below standard, meeting standard, exceeding standard in the standard setting scenario, or they may be said to represent nonmastery, intermediate mastery, and mastery of the attribute. An examinee having a higher mastery level of an attribute also possesses the lower mastery level of the attribute. Using the same example, if an examinee is exceeding the standard, he or she must have met the standard first. In general, if an attribute a has l_a mastery levels, the values of the attribute are 0, 1, … (l_a-1). Suppose a test measures *A* attributes, there are in total $\prod_{a=1}^{A} l_a$ attribute profiles in which examinees will be classified. Note that the GLMPM allows attributes on the same to have different numbers of mastery levels. For example, a test measures 3 attributes as shown in Table 2.2. Attribute 1 has three mastery levels and Attribute 2 and 3 have two mastery levels. Examinees are classified into $3 \times 2 \times 2 = 12$ possible attribute profiles, where the values for Attribute 1 are 0, 1 and 2, and the values for Attribute 2 and 3 are 0 and 1.

Though the mastery levels of the attribute can be more than two levels, the entries of Q-matrix are still dichotomous indicators, where 0 represents the item does not measure the attribute and 1 represents the item measures the attribute. The behavior of the polytomous attribute mastery levels is the probability of answering an item correctly where a higher mastery level means a larger probability of answering the item correctly. The model is again under the local independence assumption that the item response for each item is independent given an examinee's attribute profile.

The probability of a correct response for item *i* is generalized from the RUM. Suppose an examinee *e* has attribute profile $\boldsymbol{\alpha_e} = (\alpha_{e1}, \cdots, \alpha_{ea}, \cdots, \alpha_{eA})'$ and higher-order latent trait θ_e, and the *q*-vector for Item *i* is $\boldsymbol{q_i} = (q_{i1}, \cdots, q_{ea}, \cdots, q_{iA})'$, the measurement model for the polytomous RUM for Item *i* is

$$P(X_{ei} = 1|\boldsymbol{\alpha_e}, \theta_e) = \pi_i^* \prod_{a=1}^{A} \left({r_{ia}^*}^{f_{ia}(\alpha_{ea}, q_{ia})}\right) \times P_{c_i}(\theta_e), \tag{2.9}$$

where π_i^* is a product of the parameters that correspond to all attributes measured by item *i*,

$$\pi_i^* = \prod_{a=1}^{A} \pi_{ia}^{q_{ia}}, \tag{2.10}$$

where π_{ia} is between 0 and 1 and can be considered as the contribution to the probability of a correct response for item *i* by Attribute *a*. If Attribute *a* is present in $\boldsymbol{q_i}$, $\pi_{ia}^{q_{ia}} = \pi_{ia}$. The product of all the probabilities π_{ia} for attributes measured by Item *i* is defined as π_i^*, which is the probability of answering Item *i* correctly when an examinee has mastered all the required attributes. In this model, π_i^* is referred to as the slipping parameter. Assume all attributes contribute independently to the slipping parameter π_i^*, the probability of answering item *i* correctly is the product of all the slipping parameters for the attributes measured by item *i*, when examinee *e* has mastered all the required attributes. The item parameter r_{ia}^* can be considered as the penalty of the item response probability for not mastering a required attribute. It has the form of

$$r_{ia}^* = \frac{r_{ia}}{\pi_{ia}} \tag{2.11}$$

where r_{ia} is usually smaller than π_{ia} and thus the r_{ia}^{*} is a proportion of π_{ia}. Suppose Attribute a has l_a levels, where $\alpha_a = 0, 1, \cdots, (l_a - 1)$, Templin (2004) defined $f_{ia}(\alpha_{ea}, q_{ia})$ as

1) $f_{ia}(\alpha_{ea} = 0, q_{ia} = 1) = 1,$
2) $f_{ia}(\alpha_{ea} = (l_a - 1), q_{ia} = 1) = 0,$
3) $f_{ia}(\alpha_{ea} = 1, q_{ia} = 1) > f_{ia}(\alpha_{ea} = 2, q_{ia} = 1) > \cdots > f_{ia}(\alpha_{ea} = (l_a - 2), q_{ia} = 1),$
4) $f_{ia}(\alpha_{ea} = k, q_{ia} = 0) = 0.$

where 1) represents the index of the penalty is 1 and the penalty on the item response probability for item *i* is present when examinee *e* did not master or had the lowest mastery level of the required attribute a. 2) represents the index of the penalty is 0 and the penalty is absent when examinee *e* has the highest mastery level for attribute a. 3) represents the index of the penalty decreases as the intermediate mastery level increases, that is, the penalty on the item response probability also decreases as the mastery level increases. 4) represents the index is always equal to 0 when Attribute *a* is not measured by item *i* when examinee *e* possesses any mastery level ((*k+1)*th level), meaning there is no penalty on the item response probability for the attributes not required by item *i*. $f_{ia}(\alpha_{ea} = k, q_{ia} = 1)$ can be considered unknown in the polytomous RUM which brings $(l_a - 2)$ more item parameters per item. To reduce the number of item parameters, the author constrained $f_{ia}(\alpha_{ea} = k, q_{ia} = 1)$ to be equal for all items such that the number of item parameters for $f_{ia}(\alpha_{ea} = k, q_{ia} = 1)$ was reduced to $(l_a - 2)$ for the test.

$$f_{1a}(\alpha_{ea} = k, q_{ia} = 1) = f_{2a}(\alpha_{ea} = k, q_{ia} = 1) = \cdots = f_{Ia}(\alpha_{ea} = k, q_{ia} = 1),$$

$P_{c_i}(\theta_e)$ is the similar to the 1PL model in the item response theory with different definition of the item difficulty parameter:

$$P_{c_i}(\theta_e) = \frac{\exp\left(D(\theta_e+c_i)\right)}{1+\exp\left(D(\theta_e+c_i)\right)} \tag{2.12}$$

where *D* is the scaling constant 1.701; θ_e is examinee *e*'s latent ability; c_i represents the completeness of the Q-matrix, that is, whether the attributes specified by the Q-matrix can fully explain the item behavior. The range of c_i is from 0 to 3, where 0 means the attributes specified by Q-matrix cannot fully describe the change of the item response probability and $P_{c_i}(\theta_e)$ almost ranges from 0 to 1 as the θ_e ranges from -3 to 3. This means examinee *e*'s ability has a strong influence on the item response probability. When c_i equals 3, $P_{c_i}(\theta_e)$ is ranges from .5 to 1 as θ_e ranges from -3 to 3 meaning examinee *e*'s ability does not have a strong influence on the item response probability. The attributes specified in the Q-matrix for item *i* can almost fully explain the change of the item response probability when only considering the attributes.

Furthermore, the author assumed the mastery levels for an attribute is related to the proficiency space which includes a set of covariates and a higher-order latent trait. Examinee *e*'s mastery level for Attribute *a*, denoted as α_{ea} was mapped to a continuous variable $\tilde{\alpha}_{ea}$.

$$\tilde{\alpha}_{ea} = \boldsymbol{\beta_a y_e} + \lambda_a g_e + e_{ae} \tag{2.13}$$

where $\boldsymbol{y_e}$ is a vector of covariates; $\boldsymbol{\beta_a}$ is a vector of coefficients for the covariates for Attribute *a*; g_e is the higher-order latent trait for examinee *e*; λ_a is the coefficient for the latent trait for Attribute *a*; e_{ae} is the residual for examinee e and Attribute *a*. The range of g_e is (-1, 1) and e_{ae} follows an independent normal distribution $N(0{,}1-\lambda_a^2)$. Therefore, $\tilde{\alpha}_{ea}$ follows a normal distribution $N(\boldsymbol{\beta_a y_e} + \lambda_a g_e, 1-\lambda_a^2)$. To map $\tilde{\alpha}_{ea}$ to the polytomous α_{ae}, the author defined $(l_a - 2)$ cut point parameters $\kappa_{a1}, \cdots, \kappa_{a(l_a-2)}$, such that for the *k*th mastery level,

$$I(\alpha_{ea} > k) = I(\tilde{\alpha}_{ea} > \kappa_{ak}) \tag{2.14}$$

We also have

$$P(\tilde{\alpha}_{ea} > \kappa_{ak}|\boldsymbol{Y_e} = \boldsymbol{y_e}, G_e = g_e) = 1 - \Phi\left(\frac{\boldsymbol{\beta_a y_e} + \lambda_a g_e - \kappa_{ak}}{\sqrt{1-\lambda_a^2}}\right) \tag{2.15}$$

where $\Phi(\cdot)$ is the cumulative density function of the standard normal distribution. Therefore, the probability of examinee *e* having mastery level *k* for Attribute *a* is

$$P(\alpha_{ea} = k|\boldsymbol{Y_e} = \boldsymbol{y_e}, G_e = g_e) = P(\alpha_{ea} > k|\boldsymbol{Y_e} = \boldsymbol{y_e}, G_e = g_e) -$$

$$P(\alpha_{ea} = (k+1)|\boldsymbol{Y_e} = \boldsymbol{y_e}, G_e = g_e) = P(\tilde{\alpha}_{ea} > \kappa_{ak}|\boldsymbol{Y_e} = \boldsymbol{y_e}, G_e = g_e) -$$

$$P\big(\tilde{\alpha}_{ea} > \kappa_{a(k+1)}\big|\boldsymbol{Y_e} = \boldsymbol{y_e}, G_e = g_e\big) = \Phi\left(\frac{\boldsymbol{\beta_a y_e} + \lambda_a g_e - \kappa_{a(k+1)}}{\sqrt{1-\lambda_a^2}}\right) - \Phi\left(\frac{\boldsymbol{\beta_a y_e} + \lambda_a g_e - \kappa_{ak}}{\sqrt{1-\lambda_a^2}}\right) \tag{2.16}$$

To assess the effectiveness the RUM for the polytomous attributes under the GLMPM framework, Templin (2004, Chapter 4) conducted a set of simulation studies. The simulation studies evaluated two polytomous RUM under two test complexity (number of Q-matrix entries) conditions, two cognitive structure (magnitude of r_i^* parameters) conditions, and two completeness (magnitude of *c* parameters) conditions. The first polytomous RUM had the constraint that the item parameters f_{ia} were the same for each mastery level across all items, while the second model allowed them to be different. The results were evaluated by the recovery accuracy of the item parameter estimates and the classification accuracy of examinees' attribute profiles. The two models yielded similar results: all the item parameters had relatively low biases across all conditions. The estimates of *c* parameters were positively biased when the test completeness was low and negatively biased when the test completeness was high. The estimates of f_{ia} had larger biases when the cognitive structure was high. The classification accuracies for examinees' attribute profiles were lower compared to the results in the dichotomous simulation studies.

The author used the Fraction Subtraction data set to examine the two polytomous RUM. The Fraction Subtraction test measured 8 attributes with 20 items. The author assumed each attribute had three mastery levels (0, 1, and 2). The results shown the test completeness increased when using the polytomous RUM compared to dichotomous RUM. The results also indicated it was not always appropriate to use polytomous attributes because of the lack of examinees.

Ordered-Category Attribute Coding framework (OAOC)

Different from Templin's (2004) polytomous RUM that only defined the order of attribute mastery levels, Karelitz (2004) suggested to provide specific definitions of the mastery levels for each attribute, and assumed examinees' current mastery levels of the attributes measured by the test followed the sequences of the lowest levels to the highest levels. He proposed an Ordered-Category Attribute Coding (OAOC) framework to model the item response probabilities of the polytomous attribute levels.

Like the polytomous RUM, the OAOC framework defined the attributes as integers from 0 to $(l_a - 1)$, where l_a is the number of mastery levels for Attribute a. The difference is that the author assumed each attribute mastery level had a cognitive definition and examinees must have mastered knowledge or skills represented by all lower levels to achieve the higher-level proficiency. The ordered mastery levels of an attribute are a reflection of an examinee's learning process of certain knowledge, or the steps required to reach a correct response. The definition of each mastery level provides more detailed diagnostic information compared to the polytomous RUM. However, the application of the OAOC framework requires content experts to carefully design the test with clarification of each mastery level of the attributes, while the polytomous RUM has more flexibility in deciding the number of attribute mastery levels. For example, model selection criteria, such as AIC, BIC and SABIC, can be used to compare models under different attribute mastery levels and further help decide the number of mastery levels for each attribute. Suppose a test measures *A* attributes and each attribute has l_a mastery levels, the number of attribute profiles in the OAOC framework is $\prod_{a=1}^{A} l_a$.

The definition of the Q-matrix is the main difference between the OAOC framework and the polytomous RUM. In the polytomous RUM, the entries of the Q-matrix are dichotomous indicating whether the attribute is measured by the item. In the OAOC, the entries of the Q-matrix are polytomous indicating which mastery level is measured by the item. For example, assume Attribute *a* has three mastery levels and Item *i* measures Attribute *a*, the entry of the *i*th row and *a*th column q_{ia} can be 0, 1, or 2. 0 represents Item *i* does not measure Attribute *a*; 1 represents Item *i* measures the first mastery level of Attribute *a;* and 2 represents Item *i* measures the highest mastery levels of Attribute *a*. Suppose Item *i* measures the first level of Attribute *a*.Examinees who have the lowest mastery level will have a relatively low probability of answering Item *i* correctly, while examinees who have the second or the highest mastery level will have equally higher probability of answering Item *i* correctly. In the polytomous RUM, the item response probability strictly increases as the examinees' mastery levels increase. Since the OAOC framework specifies more detailed attribute mastery levels that were measured by each item, the test might need enough items to measure each attribute mastery levels to guarantee the accurate classification of examinees.

In this study, the author generalized the DINA model to the OAOC framework. Assume a test measures *A* attributes, the item response probability for examinee *e* answering Item *i* correctly is

$$\mathrm{P}(\mathrm{X}_{\mathrm{ei}} = 1|\boldsymbol{\alpha_e}) = (1 - s_i)^{\xi_{ei}} g_i^{1-\xi_{ei}}, \tag{2.17}$$

$$\xi_{\mathrm{ei}} = \prod_{a=1}^{A} I[\alpha_{ea} \geq q_{ia}],$$

where $\boldsymbol{\alpha}_{\mathbf{e}} = (\alpha_{e1}, \cdots, \alpha_{eA})$ is the attribute profile of examinee *e*. The entry $\alpha_{\mathrm{ea}} = 0, 1, \cdots, (l_a - 1)$ of $\boldsymbol{\alpha}_{\mathbf{e}}$ represents examinee *e*'s mastery level of Attribute *a*. s_i is the slipping parameter representing the probability of answering the item incorrectly when examinee *e* has mastered all the required attribute levels measured by item *i*; g_i is the guessing parameter representing the probability of answering the item correctly when examinee *e* has not mastered all the required attribute levels; ξ_{ei} is a dichotomous indicator of whether examinee *e* has mastered all the required attribute levels of item *i*. ξ_{ei} is a product of the indicator function of whether α_{ea} is larger than q_{ia}. $I[\alpha_{ea} \geq q_{ia}]$ equals 1 when α_{ea} is larger than q_{ia}, meaning examinee *e* has higher mastery level for Attribute *a* than the required attribute mastery of item *i*, otherwise, equals 0. ξ_{ei} equals 1 if and only if all the $I[\alpha_{ea} \geq q_{ia}]$ equals 1, meaning examinee *e* has mastered all the required attributes for item *i*, otherwise equals 0. When examinee *e* has mastered all the required attribute levels of item *i*, the item response probability equals $(1 - \mathrm{s_j})$ because the index of g_i equals 0. When examinee *e* has not mastered at least one required attribute of item *i*, the probability of a correct response is g_i.

The author conducted two sets of simulation studies to investigate the performance of the OAOC framework. The first set examined the stability of the model calibration under the conditions of different numbers of attributes and mastery levels, distributions of the proportions of attribute profiles, the numbers of missing entries of the Q-matrix, different patterns of the missing entries in the Q-matrix, and the sensitivity of the item-level noise parameters. The second set examined the robustness of the model when the model assumptions were violated: the estimation model had more levels than the simulation model, the estimation model had fewer levels than the simulation model, the estimation model only had two levels for all attributes with item-level noise parameters. The models were estimated by MCMC algorithm.

The results of the simulation studies shown the OAOC combined with DINA model as measurement model can accurately estimate item parameters and classify examinees under different number of attributes and mastery levels even there was item-level noise in the parameters. The OAOC framework was sensitive to the distribution of the proportions of attribute profiles. The increase of the number of missing entries in the Q-matrix decreased the accuracy of the item parameter estimation and examinees' classification. The asymmetric attributes missing entries of the Q-matrix had lower accuracy for the item parameter estimates and examinees classification than the symmetric attribute missing of the Q-matrix. The over-specification and under-specification of the attribute mastery levels would change the OAOC framework and provided incorrect classifications for examinees.

The author used the item responses from a test of the grammatical rules of a fictional language conducted at University of Illinois. The sample size was 200. The test measured three attributes, of which two attributes had four mastery levels and one attribute had three mastery levels. The test contained 40 multiple-choice items. The authors assumed the first levels of all attributes were the nonmastery level. Attribute 1 was measured by 39 items, among which 12 items measured the second level, 11 items measured the third level, and 6 items measured the fourth level. Attribute 2 was measured by 30 items, among which 8 items measured the second level, 11 items measured the third levels and 11 items measured the fourth levels. Attribute 3 was measured by 25 items, among which 14 items measured the second level and 11 items measured the third level. The results show many items had large guessing parameter estimates.

The Hierarchical DCM

A similar concept to a polytomous attribute is a linear attribute hierarchy (Templin and Bradshaw, 2014). The linear attribute hierarchy in the DCMs represents that attributes measured by a test follow a sequence to reflect examinees' learning. For example, assume a test measures three attributes (Attribute 1, 2, and 3) , a linear attribute hierarchy of the three attributes could be: Attribute 1 is the prerequisite of Attribute 2, and Attribute 2 is the prerequisite of Attribute 3. This means an examinee who is a master of Attribute 2 must be a master of Attribute 1, and an examinee who is a master of Attribute 3 must be a master of Attribute 2 and Attribute 1. Assume there are three exmainees e_1, e_2 and e_3 who are masters of Attribute 1, Attribute 2 and Attribute 3 respectfully, then Examinee e_2 is also a master of Attribute 1 and Examinee e_3 is also a master of Attribute 2 and 3. Thus, Examinee e_3 can be considered to have the highest attribute mastery level, followed by Examinee e_2, and Examinee e_1 has the lowest attribute mastery level. In this case, the three linear hierarchical attributes are equivalent to one polytomous attribute with four mastery levels. The corresponding attribute profiles for three attributes with a linear hierarchy and a polytomous attribute are presented in Table 2.4.

Since the linear attribute hierarchy is present, examinees can only be classified into one of the four attribute profiles: 1) the first attribute profile is the nonmastery group for all three attributes which correspond to the nonmastery group of the polytomous attribute; 2) the second attribute profile is the mastery for Attribute 1 and nonmastery for Attribute 2 and 3, which corresponds to the first mastery level (equals 1) for the polytomous attribute; 3) the third attribute profile is the mastery for Attribute 1 and 2 and nonmastery for Attribute 3, which corresponds to the second mastery level (equals 2) for the polytomous attribute; 4) the fourth attribute profile is the mastery for all three attributes which corresponds to the highest mastery level (equals 3) for the polytomous attribute.

Templin and Bradshaw (2014) mentioned the attribute linear hierarchy can be detected by a statistical model and proposed the hierarchical DCM (HDCM) to model the item response probability. The authors also addressed the HDCM is similar to the unidimensional DCM (UDCM) for a multicategory attribute.

Using the same example, the item response probability for Item *i* for the UDCM is

$$\log \frac{P(X_{ei}=1|\alpha_{ea})}{P(X_{ei}=0|\alpha_{ea})} = \lambda_{i0} + \lambda_{ia}\alpha_{ea} \tag{2.18}$$

where α_{ea} is the mastery levels for examinee *e* which can be equal to 0, 1, 2 or 3. λ_{i0} is the intercept which represents the log-odds of the item response probability for the nonmastery group; λ_{ia} is the main effect meaning the increase of the log-odds when examinee *e* has one higher mastery level.

The authors compared the 3-attribute HDCM and the UDCMs with 2 to 5 attribute categories using the item responses from the Examination for the Certificate of Proficiency in English (ECPE; Henson & Templin, 2007; Templin & Hoffman, 2013; Templin, Rupp, Henson, Jang, & Ahmed, 2008; and Buck & Tatsuoka, 1998). AIC, BIC and SSA BIC indicated the UDCM with 5 categories fitted the best. The authors suggested the UDCM can be used to evaluate whether a certain number of attributes follow a linear hierarchy structure. Moreover, the UDCM is also the simplest model for a unidimensional polytomous attribute.

The Polytomous Generalized DINA

Chen and de la Torre (2013) referred polytomous DCMs like the polytomous RUM (Templin, 2004) as the *data-defined* polytomous attributes and the OAOC framework as the *expert-defined* polytomous attributes. Though the OAOC provides more detailed diagnostic information for the mastery levels of each attribute, the DINA model used in the OAOC cannot comprehensively explain the relationships of the attributes. The authors generalized the measurement model of the OAOC framework to a general DCM framework called the generalized DINA (G-DINA, de la Torre, 2011). The combination of the OAOC framework and the G-DINA model is called the pG-DINA.

Since the pG-DINA is based on the OAOC framework, examinees' latent attributes values and the Q-matrix entries have the same definition as the OAOC framework. Suppose a test measures *A* attributes and Attribute *a* has l_a mastery levels and is measured by Item *i*, an examinee's mastery level for Attribute *a* and the possible values for the entry q_{ia} of the Q-matrix might be 0, 1, …, $(l_a - 1)$. The probability of examinee *e* answering item *i* correctly is

$$P(X_{ei} = 1|\boldsymbol{\alpha}_{\boldsymbol{e}}^{**}) = \delta_{i0} + \sum_{a=1}^{A_i^*} \delta_{ia}\, \alpha_{ea}^{**} + \sum_{a'>a}^{A_i^*} \sum_{a=1}^{A_i^*} \delta_{iaa'} \alpha_{ea}^{**} \alpha_{ea'}^{**} + \cdots +$$

$$\delta_{i1,\cdots,A_j^*} \prod_{a=1}^{A_i^*} \alpha_{ea}^{**}, \tag{2.19}$$

$$\alpha_{ea}^{**} = \begin{cases} 0, & if\ \alpha_{ea} < q_{ia} \\ 1, & otherwise \end{cases},$$

where $\boldsymbol{\alpha}_{\mathbf{ei}}^{**} = (\alpha_{e1}^{**}, \cdots, \alpha_{eA^*}^{**})'$ is a dichotomized vector of $\boldsymbol{\alpha}_{\boldsymbol{ei}}$, called the *collapsed attribute vector*, indicating whether examinee *e* has mastered the required mastery level of an attribute measured by item *i*. The elements of $\boldsymbol{\alpha}_{\boldsymbol{ei}}^{**}$ are only the attributes measured by item *i*. For example, if a test measures 4 attributes and item *i* measures only 2 attributes (Attribute 1 and 3), $\boldsymbol{\alpha}_{\mathbf{ei}}^{**} = (\alpha_{e1}^{**}, \alpha_{e3}^{**})'$. α_{e1}^{**} equals 0 if examinee e possesses lower mastery level than the mastery level item i measures q_{ia}, and equals 1 if examinee e possesses equal to or higher mastery level than q_{ia}.

The authors then applied $\boldsymbol{\alpha}_{\boldsymbol{ei}}^{**}$ to the G-DINA model with the identity link function of the item response probability. δ_{i0} is the intercept, representing the probability of examinee *e* guessing item *i* correctly when the examinee does not possess any required mastery levels of the attributes measured by item *i*. δ_{ia} is the main effect of Attribute *a* for item *i*, representing the increase of the probability of a correct response for item *i* when examinee *e* has mastered the required mastery level of Attribute a. $\delta_{iaa'}$ is the two-way interaction of Attribute *a* and *a'*. $\delta_{i1,\cdots,A_j^*}$ is the highest order of interaction term. The G-DINA model is an item level measurement model and the item response probability is the summation of all intercept, main effects and interactions.

The authors also proposed another polytomous attribute model, called mG-DINA, which classified examines by placing multiple cut-off points on the attribute posterior probability scale. For example, assume Attribute *a* has three mastery levels, an examinee's attribute posterior probability being classified into two mastery levels under G-DINA is between 0 and 1. The cut-off points for the three mastery levels under mG-DINA are set to be 0.40 and 0.60. An examinee with the attribute posterior probability 0.10 under G-DINA is classified into the nonmastery group under the mG-DINA; an examinee with the attribute posterior probability 0.45 under G-DINA is classified into the intermediate mastery group; and an examinee with the attribute posterior probability 0.80 under G-DINA is classified into the mastery group. The authors also compared the pG-DINA with a mG-DINA which has the same form as G-DINA model but classifies examinees' mastery levels into polytomous mastery levels based on the posterior probability of an attribute.

The authors conducted a set of simulation studies to examine the performance of the pG-DINA under two numbers of attributes, three sample sizes, two test lengths, two Q-matrices with the attribute mastery levels as 2 and 3. Instead of using the saturated pG-DINA, the author used the constrained versions, where one model contained only the intercept and main effects, called polytomous-attribute A-CDM, and the other model contained only the intercept and the highest-order of interaction, called polytomous-attribute DINA. The data was generated by the pG-DINA with the lowest item response probability .1 and the highest item response probability .90. Then they estimated the data under the pG-DINA and mG-DINA using the software OX through EM algorithm.

The results for the polytomous-attribute A-CDM and the polytomous-attribute DINA model were similar. The item parameter estimates had small biases and the standard errors. The standard errors of the A-CDM were larger than those of DINA model, and decreased as the sample size and test length increased, and the Q-matrix was less complex. The classification of examinees' mastery levels ranged from .76 to .93. The classification accuracy increased as the sample size increased and the number of mastery levels decreased. The pG-DINA had more accurate classifications than the mG-DINA.

The authors examined the pG-DINA model with a real data example. The data was collected from a proportional reasoning assessment for eighth-grade students. The test measured 4 attributes by 15 items. Two attributes were dichotomous, and another two attributes were polytomous with 3 mastery levels. 8 items were simple items that measured only one attribute and 7 items were complex items that measured more than one attribute. The sample size was

393. The results show only around 10% of examinees were classified into the intermediate mastery levels for the polytomous attributes. Results also show some issues with item parameter estimates. The item response probability for Item 1 when an examinee was nonmastery of the required attributes was .82, meaning the item was an easy item and many examinees answered this item correctly without mastering the attribute. Three items had low item response probabilities for the mastery group, meaning these items were difficult items and very few examinees answered this item correctly even they mastered the required attributes. Four items violated the assumption that the mastering more attributes can improve the item response probability, that is, these items had at least one item response probability smaller than the one correspond to mastering less attributes.

The Generalized Diagnostic Model

von Davier (2005) proposed the general diagnostic model (GDM) and generalized it for polytomous responses, called pGDM. The pGDM provides a very general formulation and as such subsumes most measurement models including item response theory models such as the Rasch, 2PL, generalized partial credit model, and the FACETS model, as well as many DCMs (von Davier, 2005). Specially, von Davier (2005) mentioned a special case of the pGDM with the polytomous attributes. The attribute definition under the pGDM is similar to the definition under the polytomous RUM, except that the values of the attributes can be any integers. This definition is more like the categorizing the latent abilities under the multidimensional IRT model (Reckase, 2009). Moreover, the author assumed the increase of the latent ability was equal between any two adjacent mastery levels.

The form of pGDM can be considered as a constraint version of the polytomous RUM with simple mathematical transformation. The difference of the GDM is in how the attribute levels are defined. More specifically, suppose Attribute *a* has l_a mastery levels, the mastery level of Attribute *a* in the GDM is $\alpha_a \in \{s_{a1}, s_{a2}, \cdots, s_{al_a}\}$. s_{al} can be defined as any increasing sequence of integers. For example, the most common scenario is that Attribute *a* has two mastery levels: mastery and nonmastery. If we define mastery as 1 and nonmastery as 0, the mastery level $\alpha_a \in \{0,1\}$. Another example of the mastery level of Attribute *a* is $\alpha_a = \{-m, -(m - 1), \cdots, -1, 0, 1, \cdots, (m - 1), m\}$. In this example, Attribute *a* has $(2m + 1)$ levels and the value for each level is symmetric around 0 ranging from $-m$ to m, where $-m$ represents the lowest mastery level (nonmastery) and m represents the highest mastery level. Haberman et al. (2008) referred the GDM with this definition of the polytomous attributes as the multidimensional item response theory with polytomous latent variables and compared the model with the multidimensional item response theory with continuous latent variables.

Besides the definition of attribute mastery levels, the definition of the Q-matrix entries in pGDM is polytomous. Suppose a test measures *A* attributes, the entries of the Q-matrix for item *i* are $\boldsymbol{q_i} = (q_{i1}, q_{i2}, \cdots, q_{iA})$. The entry of $\boldsymbol{q_i}$, for example q_{ia}, can be any non-negative integers based on the user’s definition. Usually, $\mathrm{q_{ia}} = 0$ represents Item *i* does not measure Attribute *a* and $\mathrm{q_{ia}} > 0$ represents Item *i* measures Attribute *a*. For example, if Attribute *a* has two mastery levels, $\mathrm{q_{ia}} \in \{0,1\}$, where 0 represents Item *i* does not measure Attribute *a* and 1 represents Item *i* measures Attribute *a*. If Attribute *a* has three mastery levels, one possible definition for q_{ia} is $q_{ia} \in \{0,1,2\}$, where 0 represents Item *i* does not measure Attribute *a*, 1 represents Item *i* measures the second mastery level of Attribute *a*, and 2 represents Item *i* measures the third mastery level of Attribute *a*.

In summary, the definition of the attribute mastery levels and the definition of Q-matrix entries are different under pDCM. The values of the attribute mastery levels are symmetric with respect to 0, where the numbers of negative values for mastery levels are the same as the numbers of positive values for mastery levels. The values of Q-matrix entries under pDCM are polytomous integers ranging from 0 to the highest mastery levels. Although the definitions of attribute mastery levels and Q-matrix entries are different, the numbers of the values used in the two definitions are the same and equal to the number of attribute mastery levels.

For examinee *e* with attribute profile $\boldsymbol{\alpha}_{\mathbf{e}}$, the probability of answering Item *i* correctly is

$$\log\frac{P(X_{ei}=1|\boldsymbol{\alpha}_{\boldsymbol{e}})}{P(X_{ei}=0|\boldsymbol{\alpha}_{\boldsymbol{e}})} = \lambda_{i0} + \boldsymbol{\lambda}_{\boldsymbol{i}}^{\boldsymbol{T}}\boldsymbol{h}(\boldsymbol{\alpha}_{\boldsymbol{e}}, \boldsymbol{q}_{\boldsymbol{i}}), \tag{2.20}$$

where λ_{i0} is the intercept of the log odds of the item response probability. $\boldsymbol{\lambda}_{\boldsymbol{i}}^{\boldsymbol{T}}$ is a vector of other item parameters representing the contribution by each attribute measured by the test. $\boldsymbol{h}(\boldsymbol{\alpha}_{\boldsymbol{e}}, \boldsymbol{q}_{\boldsymbol{i}}) = \left(h_1(\boldsymbol{\alpha}_{\boldsymbol{e}}, \boldsymbol{q}_{\boldsymbol{i}}), \cdots, h_A(\boldsymbol{\alpha}_{\boldsymbol{e}}, \boldsymbol{q}_{\boldsymbol{i}})\right)$ is a vector of *A* functions, where each function represents a specific relationship between examinee *e*'s attribute profile and the Q-matrix entries $\boldsymbol{q}_{\boldsymbol{i}}$. For example, if a Q-matric entry q_{ia} is defined as $q_{ia} \in \{0,1\}$, where 0 represents Item *i* does not measure Attribute *a*, and 1 represents Item *i* measures Attribute *a.* Attribute *a* measures l_a levels, where $\alpha_{ea} \in \{0, 1, \cdots, (l_a - 1)\}$. We define $h_a(\boldsymbol{\alpha}_{\boldsymbol{e}}, \boldsymbol{q}_{\boldsymbol{i}}) = \alpha_{ea}q_{ia}$, then

$$h_a(\boldsymbol{\alpha}_{\boldsymbol{e}}, \boldsymbol{q}_{\boldsymbol{i}}) = \begin{cases} \alpha_{ea} & when\ q_{ia} = 1 \\ 0 & when\ q_{ia} = 0 \end{cases}.$$

Suppose Item *i* measures Attribute a_1 and a_2, the item response probability for Item *i* is

$$\log\frac{P(X_{ei}=1|\boldsymbol{\alpha}_{\boldsymbol{e}})}{P(X_{ei}=0|\boldsymbol{\alpha}_{\boldsymbol{e}})} = \lambda_{i0} + \lambda_{i1}\alpha_{ea_1} + \lambda_{i2}\alpha_{ea_2} \tag{2.21}$$

where λ_{i1} is the coefficient for Attribute a_1, λ_{i2} is the coefficient for Attribute a_2. The log-odds of $P(X_{ei} = 1|\boldsymbol{\alpha}_{\boldsymbol{e}})$ is the linear combination of the mastery levels for Attribute a_1 and a_2. Since the polytomous attribute GDM is a special case of the pGDM, the author did not provide any simulation study or empirical study to examine this model under various conditions.

Other Polytomous Q-matrix Design Theories

Besides the studies about the measurement models of polytomous attributes, some research (Ding et al., 2016) focus on the design of the Polytomous Q-matrix. Since the increase of mastery levels will largely increase the number of attribute profiles, it may require a much longer and less complex test design to yield the same classification accuracy. Ding et al. (2016) proposed a theorem about the minimum requirement of the polytomous Q-matrix design when attribute hierarchical structure was present. The minimum requirement of the polytomous Q-matrix design will guarantee to discriminate examinees from different attribute profiles to the lowest extent. The more items are added to the test, the more powerful the test can classify examinees.

Discussion

In this chapter, we reviewed the current studies about the polytomous DCMs. We summarized the definition of the attributes, the test design, the measurement model, the simulation studies and the applications for each model, respectively. Table 2.4 presents the comparison about the attribute values, Q-matrix entries, the structural model, and the measurement model of the four polytomous DCMs. The models can provide more detailed diagnostic information than the traditional dichotomous DCMs by classifying examinees' mastery levels into more than two levels. To review the four existing polytomous DCMs, this chapter follows a sequence of introducing: *1) the definition of polytomous attributes; 2) the design of Q-matrix; 3) the structural model; 4) the measurement model.*

The Definition of Polytomous Attributes

For the definition of polytomous mastery levels of an attribute, the polytomous RUM and the OAOC framework with either DINA or G-DINA as the measurement model defined the mastery levels as nonnegative integers from 0 to the highest mastery level, where the mastery level increases as the value increases. They all have the assumption that an examinee who has a higher mastery level of an attribute also masters the lower levels. Specifically, the polytomous RUM uses the latent attribute values to indicate the order of the mastery levels, while the OAOC framework requires the specific definition for every mastery level of an attribute.

The diagnostic feedback the models provide is also different: the polytomous RUM specifies the ordered mastery levels examinees possesses; since the OAOC framework needs the definition of each attribute mastery level prior to the test administration, the attribute classification gives specific feedback corresponding to the definition of each attribute mastery level. The GDM has a vaguer definition of the attributes which can be either the nonnegative values like the polytomous RUM and the OAOC, or symmetric discrete values with respect to 0 like a polytomous multidimensional item response theory model. von Davior (2005) did not mention whether the mastery levels of an attribute had the unique definition like the OAOC. The definition of the polytomous attribute requires the collaborative work between psychometricians and content experts for each attribute mastery levels.

The Design of Q-matrix

Since the polytomous RUM does not define attribute mastery levels, the entries of the Q-matrix are dichotomous indicating whether the attribute is measured by an item, while the OAOC framework requires polytomous entries indicating which mastery level of an attribute the item measures. Again, the GDM provides a general definition of the entries of the Q-matrix. The Q-matrix entries can be either dichotomous as the polytomous RUM or polytomous like the OAOC, or even any value ranging from negative infinity to positive infinity which researchers might find meaningful.

The Structural Model

Among the four major studies about the polytomous DCMs, only the study of the polytomous RUM provided detailed explanation of the structural model to describe the correlations among attributes. The structural model was constructed based on the loglinear regression model with the main effects and interactions of the attributes and combined the possible higher-order latent trait and other covariates. The loglinear formation allows DCM users to have more flexibilities. Moreover, adding the higher-order latent trait and covariates also help to classify examinees more accurately to the mastery levels. Though the OAOC and GDM framework did not combine the measurement model with the structural model, any structural model mentioned in the literature can be applied to the OAOC and GDM framework.

The Measurement Model

The four polytomous DCM models also used different measurement models. The polytomous RUM is a multiplication of the attribute effects and the ability effect on the item. The OAOC framework was applied to the DINA and G-DINA model. Both models required to transfer polytomous attributes to the dichotomous attributes based on if an examinee has possessed the required mastery levels of the attribute. Therefore, examinees who have lower mastery levels of an attribute than the required mastery level have the same probability of providing a correct response. The GDM is the linear combination of all the main effects of the attributes measured by the item and does not include any interaction terms.

Limitations of Real Data Applications

DCMs have gained increasing attention within past two decades. Researchers not only proposed new models or methods to deepen the theoretical basis of DCMs, but also applied DCMs into real world assessments to provide diagnostic feedback to examinees. However, more efforts are needed in facilitating the use of DCMs. Among the 49 empirical studies of DCMs since 2009, only 6 assessments were designed for DCMs and provided detailed test development procedures (Kunina-Habenicht, Rupp, and Wilhelm, 2009; Kim and Kim, 2013; Liu et al., 2013; Bradshaw et al., 2014; Chiu, Köhn, and Wu, 2016; You, et al., 2018). The remaining studies retrofitted assessments built under IRT or CTT to DCMs. Such studies might have the following limitations:

1) the attributes were highly correlated since the assessments were developed for the unidimensional IRT (Lee and Sawaki, 2009; Chen, Ferron, Thompson, Gorin, and Tatsuoka, 2010; Wang and Gierl, 2011; Choi, Lee, and Park, 2015; Skaggs, Wilkins, and Hein, 2016 & 2017; Liu, Huggins-Manley, and Bulut, 2018);
2) the test length is not enough to provide accurate and reliable attribute classification (Lee, Park, and Taylan, 2011; Briggs and Circi, 2017);
3) the sample size is not enough to provide accurate item parameter estimation (Im and Yin, 2009).

Moreover, only 16 empirical studies used the general DCM framework including the LCDM, the G-DINA, the GDM to estimate item responses. The use of submodels might fail to capture item-level attribute behaviors and misclassify examinees into different mastery groups (Bradshaw and Templin, 2014). Only one empirical study (Bradshaw et al., 2014) used the structural model to model the correlation and base rates of attributes. All empirical assessments measured dichotomous attributes.

In this dissertation, we focus on proposing new DCMs for polytomous attributes and providing guidance to practitioners about the test design and sample size requirements, as well as an illustration of applying the new DCMs to a diagnostic assessment to provide more detailed feedback to examinees.

Table 2.1

Sample Q-matrix from DTMR Assessment

ItemNo.	RU	PI	APP	MC	Attribute Measured per Item
1	1	0	0	0	1
2	0	0	1	0	1
3	0	1	0	0	1
4	1	0	0	0	1
5	1	0	0	0	1
6	0	1	0	0	1
7	1	0	0	0	1
8	0	0	1	0	1
9	0	0	1	0	1
10	0	0	1	0	1
11	0	0	1	0	1
12	1	0	0	0	1
13	0	0	0	1	1
14	1	0	0	1	2
15	1	0	0	1	2
16	1	0	0	0	1
17	1	0	0	0	1
18	0	1	0	1	2
19	1	1	0	0	2
20	0	1	0	1	2
21	0	1	0	0	1
22	0	1	0	0	1
23	1	0	0	0	1
24	0	1	0	0	1
25	1	1	0	0	2
26	1	0	0	0	1
27	1	1	0	0	2
Total	14	10	5	5	

Table 2.2

Attribute Profiles When Attribute 1 Has Three Mastery Levels and Attribute 2 and 3 Have Two Mastery Levels

	AP1	AP2	AP3	AP4	AP5	AP6	AP7	AP8	AP9	AP10	AP11	AP12
Attribute 1	0	1	2	0	1	2	0	1	2	0	1	2
Attribute 2	0	0	0	1	1	1	0	0	0	1	1	1
Attribute 3	0	0	0	0	0	0	1	1	1	1	1	1

Table 2.3

Attribute Profiles for the HDCM and the UDCM

Attribute Profiles for the HDCM			Attribute Profiles for the UDCM
Attribute 1	Attribute 2	Attribute 3	The 4-category Attribute
0	0	0	0
1	0	0	1
1	1	0	2
1	1	1	3

Table 2.4

Comparisons of the Current Polytomous DCMs

Model	Attribute Values	Q-matrix entries	Structural Model	Measurement Model
pRUM	Polytomous (nonnegative)	Dichotomous	Log-linear	RUM
Multicategory UDCM	Polytomous (nonnegative)	Dichotomous	NA	LCDM
OAOC-DINA	Polytomous (nonnegative)	Polytomous	NA	DINA
OAOC-pG-DINA	Polytomous (nonnegative)	Polytomous	NA	G-DINA
GDM	Polytomous	Dichotomous/ Polytomous	NA	GDM

CHAPTER 3

THE POLYTOMOUS-ATTRIBUTE DIAGNOSTIC CLASSIFICATION MODEL

The previous chapter described the existing DCMs for polytomous attributes. The chapter compares the polytomous-attribute DCMs and classified these models under two different test designs. This chapter generalizes the idea of the polytomous-attribute DCM from Templin (2004) and Templin and Bradshaw (2013) and proposes a new and general DCM to measure polytomous-attribute attribute mastery levels. This chapter starts to introduce the proposed polytomous-attribute DCM (PDCM) by presenting the measurement and structural models. Chapter 3 then relates the proposed model to the constrained version of the PDCM (cPDCM) which has been used in prior research. Compared to the PDCM, the cPDCM may offer a more parsimonious solution to modeling polytomous attributes when the model-data fit is adequate.

The Polytomous-attribute Diagnostic Classification Model (PDCM)

In the PDCM, we specify each attribute as an ordered polytomous latent variable, where each latent variable value represents a mastery level for the attribute. The mastery levels are ordered as the polytomous values, where lower mastery levels are defined by smaller integers, and higher mastery levels are defined by larger consecutive non-negative integers. In this section, we propose a general PDCM to explain the probability of a correct response for an item, referred to as item response probability, using the polytomous attributes. Similar to the general dichotomous DCMs, the item response probability increases as the number of required attributes an examinee has mastered increases, the difference is that the PDCM also yielded higher item response probability if the examinee attribute has a higher mastery level of the required attribute. In this section, we explain the general form of the PDCM in detail by two major components: the measurement model that models the relationship between item responses and attributes, and the structural model that illuminates the attribute correlation and attribute base rates.

Definition of Polytomous Attributes

In this study, we generalized the attribute mastery levels from dichotomous to polytomous, where we defined one or more intermediate mastery levels between non-mastery and mastery. If Attribute a has l_a mastery levels, we name these levels as the non-mastery level, the first intermediate mastery level, the second intermediate mastery level, …, the $(l_a - 2)$th intermediate mastery level, and the mastery level ($(l_a - 1)$th mastery level). Mapping these levels to positive integers, we define the latent variable α_a for attribute a with numbers $0, 1, \cdots, (l_a - 1)$, where $\alpha_a = 0$ is the non-mastery level for attribute a, $\alpha_a = 1$ is the first intermediate mastery level, $\alpha_a = 2$ is the second intermediate mastery level, $\alpha_a = x$ is the xth intermediate mastery level, and $\alpha_a = (l_a - 1)$ is the mastery level.

This definition indicates that examinees who belong to the higher mastery level of Attribute a have better understanding of the knowledge or skills associated with Attribute a than those who belong to a lower mastery level. However, the same difference between any two consecutive mastery levels values of attribute a does not mean the same difference in understanding Attribute a. For example, suppose Examinee $e_1, e_2,$ and e_3 belong to the non-mastery group, the first intermediate mastery level, and the second intermediate mastery level, respectively. The difference of the understanding of Attribute a is not necessarily the same between Examinee e_1 and e_2 as it is between Examinee e_2 and e_3.

The PCDM allows different polytomous attributes on the same assessments to have a different number of mastery levels. For example, for an assessment that measures three attributes, Attribute 1 could have two mastery levels (non-mastery and mastery), Attribute 2 could have three mastery levels (non-mastery, intermediate mastery, mastery), and Attribute 3 could have four mastery levels (non-mastery, first intermediate mastery, second intermediate mastery, mastery). In this case, examinees will be classified into the non-mastery or the mastery groups for Attribute 1; the non-mastery, the intermediate mastery, or the mastery groups for Attribute 2; and the non-mastery, the first intermediate mastery, the second intermediate mastery, or the mastery groups for Attribute 3. In total, examinees on this test would be classified into $2 \times 3 \times 4 = 24$ groups, while in a typical DCMs with dichotomous attributes, examinees on this test would be classified into $2 \times 2 \times 2 = 8$ groups. In general, if a test measures *A* attributes, the total number of possible attribute profiles is $\prod_{a=1}^{A} l_a$, where l_a is the number of mastery levels for Attribute *a*. When all the attributes have two mastery levels, meaning $l_a = 2$, the PDCM is the same as the LCDM.

The PDCM Measurement Model

As mentioned in the previous section, suppose Attribute *a* has l_a levels, the latent attribute variable α_{ea} has the value $0, 1, \cdots, (l_a - 1)$, representing examinee *e*'s mastery level for Attribute *a* with possible values being level 0 (non-mastery), level 1 (the first intermediate mastery), …, level $(l_a - 1)$ (mastery). For an item measuring Attribute *a*, the PDCM allows the item response probability to monotonically increase as an examinee's mastery level increases. For example, if an examinee has the first intermediate mastery level, his/her probability of answering this item correctly is equal to or higher than another examinee who has the non-mastery level. To achieve this flexibility of the PDCM, we utilize dummy variables to represent attribute levels: for the latent attribute variable α_{ea}, we define $(l_a - 1)$ dummy variables $\alpha_{ea}^1, \alpha_{ea}^2, \cdots, \alpha_{ea}^{(l_a-1)}$. Table 3.1 is an example of the dummy coding for each mastery level when Attribute *a* has three mastery levels $l_a = 3$.

In this example, α_{ea} is defined as 0 (non-mastery), 1 (intermediate mastery), or 2 (mastery). α_{ea}^1 and α_{ea}^2 are two dummy variables for α_{ea}, where α_{ea}^1 represents whether examinee *e* has reached the intermediate mastery level of Attribute a, and α_{ea}^2 represents whether examinee *e* has reached the mastery level of Attribute a. Any examinee cannot reach the mastery level without reaching the intermediate mastery level, meaning α_{ea}^2 cannot equal 1 with α_{ea}^1 equaling 0. When examinee *e* has the non-mastery level ($\alpha_{ea} = 0$), the values of the dummy variables are $\alpha_{ea}^1 = 0$ and $\alpha_{ea}^2 = 0$, meaning examinee *e* has not reached the intermediate mastery level yet. When examinee *e* has the intermediate mastery level ($\alpha_{ea} = 1$), the values of the dummy variables are $\alpha_{ea}^1 = 1$ and $\alpha_{ea}^2 = 0$, meaning examinee *e* has reached the intermediate mastery level, yet has not reached the mastery level. When examinee *e* has the mastery level ($\alpha_{ea} = 2$), $\alpha_{ea}^1 = \alpha_{ea}^2 = 1$ because examinee *e* has reached both the intermediate mastery level and the mastery level.

PDCM measurement model for an example item. In the following paragraphs, we introduce the measurement model for the PDCM. Suppose Item *i* measures Attribute 1 and 2, and both attributes have 3 levels. The three mastery levels for both attributes are referred to as non-mastery, intermediate mastery, and mastery. The probability for examinee *e* answering Item *i* correctly is:

$$\log \frac{P(X_{ei} = 1|\boldsymbol{\alpha}_e)}{P(X_{ei} = 0|\boldsymbol{\alpha}_e)} = \lambda_{i,0} + \lambda_{i,1,(1)}^1 \alpha_{e1}^1 + \lambda_{i,1,(1)}^2 \alpha_{e1}^2 + \lambda_{i,1,(2)}^1 \alpha_{e2}^1 + \lambda_{i,1,(2)}^2 \alpha_{e2}^2 + \lambda_{i,2,(12)}^{11} \alpha_{e1}^1 \alpha_{e2}^1$$

$$+\lambda_{i,2,(12)}^{21}\alpha_{e1}^{2}\alpha_{e2}^{1}+\lambda_{i,2,(12)}^{12}\alpha_{e1}^{1}\alpha_{e2}^{2}+\lambda_{i,2,(12)}^{22}\alpha_{e1}^{2}\alpha_{e2}^{2} \tag{3.1}$$

Where α_{e1}^{1} and α_{e1}^{2} are the dummy variables for Attribute 1; α_{e2}^{1} and α_{e2}^{2} are the dummy variables for Attribute 2; $\lambda_{i,0}$ is the intercept and the log-odds of the item response for the complete non-mastery group ($\alpha_{e1}^{1}=\alpha_{e1}^{2}=\alpha_{e2}^{1}=\alpha_{e2}^{2}=0$); $\lambda_{i,1,(1)}^{1}$ is the main effect for the intermediate mastery level of Attribute 1 ($\alpha_{e1}^{1}=1,\alpha_{e1}^{2}=0$) which represents the increase of the item response probability when examinee *e* has reached the intermediate mastery level of Attribute 1; and $\lambda_{i,1,(2)}^{1}$ is the main effect for the intermediate mastery level of Attribute 2 ($\alpha_{e2}^{1}=1,\alpha_{e2}^{2}=0$) which represents the increase of the item response probability when examinee *e* has reached the intermediate mastery level of Attribute 2; $\lambda_{i,1,(1)}^{2}$ is the main effect for the mastery level of Attribute 1 ($\alpha_{e1}^{1}=1,\alpha_{e1}^{2}=1$) which represents the additional increase of the item response probability when examinee *e* has reached the mastery level of Attribute 1; and $\lambda_{i,1,(2)}^{2}$ is the main effect for the mastery level of Attribute 2 ($\alpha_{e2}^{1}=1,\alpha_{e2}^{2}=1$) which represents the additional increase of the item response probability when examinee *e* has reached the mastery level of Attribute 2; $\lambda_{i,2,(12)}^{11}$ is the interaction for the intermediate mastery levels of Attribute 1 and Attribute 2; $\lambda_{i,2,(12)}^{21}$ is the interaction for the mastery level of Attribute 1 and the intermediate level of Attribute 2; $\lambda_{i,2,(12)}^{12}$ is the interaction for the intermediate level of Attribute 1 and the mastery level of Attribute 2; $\lambda_{i,2,(12)}^{22}$ is the interaction for the mastery levels of Attribute 1 and Attribute 2. Interaction terms represent the change in the item response probability due to interactions among different attribute levels.

Table 3.2 shows the summary of the measurement model of Item *i*. The first two columns are the mastery levels of Attribute 1 and 2; column 3 to column 6 are the dummy variables for each attribute profile; and the last column shows corresponding log-odds of the item response after conditioning on the attribute profile. For an item like Item i that measures two attributes where each attribute has three levels, there are 9 log-odds values for the item response, meaning there are up to 9 unique item response probabilities for this item.

PDCM monotonicity constraints for example item. Because of the monotonic property of the PDCM, the item response probability increases as the mastery level for a required attribute increases. For example, the item response probability for an examinee with the intermediate mastery level for Attribute 1 is higher than the probability for an examinee with the non-mastery level for Attribute 1. Since the log-odds of the item response does not change the monotonic property of a function, the PDCM has the following constraint for the main effect of the intermediate mastery level for Attribute 1:

$$\lambda_{i,0} + \lambda_{i,1,(1)}^{1} > \lambda_{i,0},$$

which is $\lambda_{i,1,(1)}^{1} > 0$. Similarly, the constraint for the main effect of the intermediate mastery level for Attribute 2 is $\lambda_{i,1,(2)}^{1} > 0$.

The same rule applies to the main effects for the mastery levels for both Attribute 1 and 2; that is, the log odds of the item response for an examinee with the mastery level for Attribute 1 or 2 is larger than the log odds of the item response for an examinee with the intermediate mastery level for Attribute 1 or 2. On the log-odds scale, the constraint is

$$\lambda_{i,1,(1)}^{2} > 0, \qquad \lambda_{i,1,(2)}^{2} > 0.$$

In the PDCM, the interaction terms are present when an examinee possesses the intermediate mastery levels or higher mastery levels for Attribute 1 and 2. For example, if an examinee has the intermediate mastery levels for both attributes, the log-odds of the item response is the sum of the main effects for the intermediate levels of Attribute 1, $\lambda_{i,1,(1)}^{1}$, and Attribute 2, $\lambda_{i,1,(2)}^{1}$, and the interaction for the intermediate levels of Attribute 1 and 2, $\lambda_{i,2,(12)}^{11}$. Therefore, the constraint becomes the log-odds of the item response for an examinee with the intermediate levels for Attribute 1 and 2 is larger than the log-odds of the item response for an examinee only possesses the intermediate level for only one of the attributes (either Attribute 1 or 2), and the non-mastery level for the other attribute. Therefore, we have

$$\lambda_{i,0} + \lambda_{i,1,(1)}^{1} + \lambda_{i,1,(2)}^{1} + \lambda_{i,2,(12)}^{11} > \lambda_{i,0} + \lambda_{i,1,(1)}^{1},$$

$$\lambda_{i,0} + \lambda_{i,1,(1)}^{1} + \lambda_{i,1,(2)}^{1} + \lambda_{i,2,(12)}^{11} > \lambda_{i,0} + \lambda_{i,1,(2)}^{1}.$$

Using algebraic manipulations, we get

$$\lambda_{i,2,(12)}^{11} > -\lambda_{i,1,(1)}^{1}, \lambda_{i,2,(12)}^{11} > -\lambda_{i,1,(2)}^{1}.$$

Applying the same rule to the other interaction terms, we have

$$\lambda_{i,2,(12)}^{12} > -\lambda_{i,1,(2)}^{2}, \lambda_{i,2,(12)}^{21} > -\lambda_{i,1,(1)}^{2},$$

$$\lambda_{i,2,(12)}^{22} > -\lambda_{i,2,(12)}^{12} - \lambda_{i,1,(2)}^{2}, \lambda_{i,2,(12)}^{22} > -\lambda_{i,2,(21)}^{21} - \lambda_{i,1,(1)}^{2}.$$

General form of the PDCM measurement model. More generally, suppose a test measures *A* attributes, the general form of the PDCM item response function when examinee *e* with attribute profile $\boldsymbol{\alpha}_e$ responds to item *i* is:

$$\log\frac{P(X_{ei} = 1|\boldsymbol{\alpha}_e)}{P(X_{ei} = 0|\boldsymbol{\alpha}_e)} = \lambda_{i,0} + \sum_{a=1}^{A}\sum_{l=1}^{l_a-1}\lambda_{i,1,(a)}^{l}\alpha_{ea}^{l}q_{ia} +$$

$$\sum_{a=1}^{A-1}\sum_{l=1}^{l_a-1}\sum_{a'=a+1}^{A}\sum_{l'=1}^{l_{a'}-1}\lambda_{i,2,(a,a')}^{ll'}\alpha_{ea}^{l}\alpha_{ea'}^{l'}q_{ia}q_{ia'} + \cdots \qquad (3.2)$$

where α_{ea}^{l} is the dummy variable for the level *l* of attribute a; q_{ia} represents whether Item *i* measures Attribute a, where 1 indicates Item *i* measures Attribute a, and 0 indicates the opposite. The main effects or the interaction terms are present in the item response function only when examinee *e* has reached the corresponding mastery levels *and* the attribute(s) are measured by Item *i*. $\lambda_{i,1,(a)}^{l}$ is the main effect for the level *l* of attribute a, and $\lambda_{i,2,(a,a')}^{ll'}$ is the two-way interaction for the level *l* of attribute *a* and the level l' of attribute a'. The ellipsis represents the summation of the possible three-way or higher-order interactions.

The saturated PDCM is the summation of the intercept, the main effects and interactions for all the dummy attribute variables for the polytomous attributes. Nested versions of the PDCM may be specified by reducing the item parameter space for the PDCM. This may be empirically-driven or theoretically-driven. When terms in the PDCM are empirically not statistically significantly different from 0, they may be removed from the PDCM for parsimony. In contrast, terms may be set to 0 prior to an analysis to form sub-models of the PDCM that are analogous to sub-models of the LCDM. For example, if all interaction terms are set to 0 in the LCDM, the compensatory Reparameterized Unified Model (C-RUM; Rupp, Templin, & Henson, 2010) is formed. Sub-models of the PDCM have fewer parameters and would be less complex to estimate. The caution for sub-models, however, is they should only be specified after establishing appropriate model-data fit, else the attribute-item relationships may be misrepresented. Sub-models of the PDCM are not the focus of the present study; we use the saturated form of the PDCM throughout.

The PDCM Structural Model

In this section, we generalized the log-linear structural model in Chapter 2 for polytomous attributes. Different from the dichotomous DCMs where the attributes follow

Bernoulli distributions, the polytomous attributes follow categorical distributions. For example, suppose Attribute *a* has 3 mastery levels ($\alpha_a = 0, 1, 2$), referred to as non-mastery, intermediate mastery, and mastery. The distribution of α_a is a categorical distribution with probabilities of $p_0 = P(\alpha_a = 0), p_1 = P(\alpha_a = 1)$ and $p_2 = P(\alpha_a = 2)$, where $p_0 + p_1 + p_2 = 1$. We call $p_0, p_1,$ and p_2 the *base rates* for Attribute *a*. Each attribute has a base rate at every mastery level which represents the proportion of examinees who have mastery at that level for the attribute.

Structural model for an example test. Like the measurement model of the PDCM, we again used dummy latent attribute variables for the structural model. Suppose a test measures two attributes (Attribute 1 and 2) and each attribute has three mastery levels. Examinees who take the test will be classified into one of the nine mastery groups based on their item responses for the test. For all examinees, we defined the proportion of being classified into a mastery group as ν_c ($c = 1,2, \cdots ,9$). The sum of ν_c equals 1 ($\sum_{c=1}^{9} \nu_c = 1$) . Like we mentioned in the Chapter 2 about the structural model for dichotomous attributes, we treat the last attribute profile which has the mastery level for all of the attributes being measured by the test as the reference group. To represent the attribute profiles in the structural model, we use dummy variables analogous to those used in the measurement model. Table 3.3 shows the 9 attribute profiles, or mastery groups, and the corresponding dummy variables.

Using these dummy variables, the structural model for polytomous attributes contains the main effects for all dummy variables and the interactions for all the combinations of the dummy variables. The proportion of examinees having attribute profile *c* is given by:

$$\mu_c = \log \frac{\nu_c}{\nu_9} = \gamma_0 + \gamma_{1,(1)}^{1} \alpha_{c1}^{1} + \gamma_{1,(1)}^{2} \alpha_{c1}^{2} + \gamma_{1,(2)}^{1} \alpha_{c2}^{1} + \gamma_{1,(2)}^{2} \alpha_{c2}^{2} +$$

$$\gamma_{2,(12)}^{11} \alpha_{c1}^{1} \alpha_{c2}^{1} + \gamma_{2,(12)}^{12} \alpha_{c1}^{1} \alpha_{c2}^{2} + \gamma_{2,(12)}^{21} \alpha_{c1}^{2} \alpha_{c2}^{1} + \gamma_{2,(12)}^{22} \alpha_{c1}^{2} \alpha_{c2}^{2} \qquad (3.3)$$

where μ_c is the natural log of the ratio of ν_c and ν_9; γ_0 is the intercept; $\gamma^1_{1,(1)}$ and $\gamma^2_{1,(1)}$ are the main effects for the first and second mastery levels of Attribute 1; $\gamma^1_{1,(2)}$ and $\gamma^2_{1,(2)}$ are the main effects for the first and second mastery levels of Attribute 2; $\gamma^{11}_{2,(12)}$ is the two-way interaction for the first mastery levels of Attribute 1 and 2; $\gamma^{12}_{2,(12)}$ is the two-way interaction for the first mastery level of Attribute 1 and the second mastery level of Attribute 2; $\gamma^{21}_{2,(12)}$ is the two-way interaction for the second mastery level of Attribute 1 and the first mastery level of Attribute 2; $\gamma^{22}_{2,(12)}$ is the two-way interaction for the second mastery levels for Attribute 1 and Attribute 2. The main effects and interactions are present in the equation when the corresponding attribute profile has equal or higher mastery levels for the required attributes.

To identify the model, the last attribute profile, the reference group, is fixed to equal 0:

$$\mathrm{\mu}_9 = \log\frac{\nu_9}{\nu_9} = 0 = \gamma_0\gamma^1_{1,(1)} + \gamma^2_{1,(1)} + \gamma^1_{1,(2)} + \gamma^2_{1,(2)} +$$

$$\gamma^{11}_{2,(12)} + \gamma^{12}_{2,(12)} + \gamma^{21}_{2,(12)} + \gamma^{22}_{2,(12)} \tag{3.4}$$

Thus, we constrain the intercept to be equal to the negative of the sum of the remaining terms in the structural model:

$$\mathrm{\gamma}_0 = -\left(\gamma^1_{1,(1)} + \gamma^2_{1,(1)} + \gamma^1_{1,(2)} + \gamma^2_{1,(2)} + \gamma^{11}_{2,(12)} + \gamma^{12}_{2,(12)} + \gamma^{21}_{2,(12)} + \gamma^{22}_{2,(12)}\right) \tag{3.5}$$

Because the sum across $\mathrm{\mu}_\mathrm{c}$ terms equals 1, the proportion of examinees having each attribute pattern, $\mathrm{\nu}_\mathrm{c}$, can be expressed as a function of $\mathrm{\mu}_\mathrm{c}$:

$$\nu_c = \frac{\exp(\mu_c)}{\sum_{c'=1}^{9}\exp(\mu_{c'})} \tag{3.6}$$

General form of the PDCM structural model. For a test that measures *A* attributes, each attribute can have different mastery levels. Suppose Attribute a has l_a levels, the number of mastery groups is $\prod_{a=1}^{A} l_a$ and the number of structural components is $(\prod_{a=1}^{A} l_a) - 1$. The general form of the saturated structural model is:

$$\mu_c = \gamma_0 + \sum_{a=1}^{A} \sum_{l=1}^{l_a-1} \gamma_{1,(a)}^{l} \alpha_a^l + \sum_{a=1}^{A-1} \sum_{l=1}^{l_a-1} \sum_{a'=a+1}^{A} \sum_{l'=1}^{l_{a'}-1} \lambda_{2,(a,a')}^{ll'} \alpha_a^l \alpha_{a'}^{l'} + \cdots \tag{3.7}$$

Where γ_0 is the intercept; $\gamma_{1,(a)}^{l}$ is the main effect for the *l*th level of Attribute a; $\lambda_{2,(a,a')}^{ll'}$ is the two-way interaction for the *l*th level of Attribute a and l'th level of Attribute a'; and the ellipses represent the higher-order interaction terms for different levels of different attributes. Again, we have the constraint for the structural model parameters:

$$\gamma_0 = -\left(\sum_{a=1}^{A} \sum_{l=1}^{l_a-1} \gamma_{1,(a)}^{l} + \sum_{a=1}^{A-1} \sum_{l=1}^{l_a-1} \sum_{a'=a+1}^{A} \sum_{l'=1}^{l_{a'}-1} \gamma_{2,(a,a')}^{ll'} + \cdots \right) \tag{3.8}$$

The membership proportion for mastery group *c* is

$$\nu_c = \frac{\exp(\mu_c)}{\sum_{c'=1}^{\prod_{a=1}^{A} l_a} \exp(\mu_{c'})} \tag{3.9}$$

Once ν_c is estimated, we can further calculate the base rate for each attribute at each level. For example, for Attribute a, the base rate for the *l*th level p_{al} is

$$p_{al} = \sum_{c \in \{\boldsymbol{\alpha}_c | \alpha_{ca} = l\}} \nu_c \tag{3.10}$$

which is the sum of all the membership proportions of attribute profiles that measure the level *l* of Attribute a. Since p_{al} is the marginal membership proportion of Attribute a, the sum of p_{al} across all levels for Attribute *a* equals 1.

The Constrained Polytomous-attribute Diagnostic Classification Model (cPDCM)

The constrained PDCM (cPDCM) is more parsimonious version of the PDCM. Parsimony is achieved by making assumptions about the magnitude of log-odds increase of correct response between mastery levels. The cPDCM uses ordinal values for attribute mastery levels as latent predictors; the result of ordinal-valued attribute level is that the main effects are constrained to be the same for each attribute level, and the interaction terms are constrained to a specific pattern. As a result, the cPDCM has the same number of the item parameters as the dichotomous LDCM, which is considerably less complex than the PDCM. Since the cPDCM requires fewer item and structural parameters, if the data fit the cPDCM model, we expect we would need a shorter test length and a smaller sample size to achieve the same level of classification accuracy as the PDCM.

Templin and Bradshaw (2014) first proposed the model under the unidimensional DCM (UDCM) for multi-category attributes. We generalized the idea to a general multidimensional polytomous-attribute DCM in this section. von Davier (2005) also demonstrated the GDM could classify examinees into polytomous mastery levels. The GDM is specified in a sufficiently broad way to share similarities with the cPDCM. The attribute levels of the GDM were defined loosely as any value. For the cPDCM, we define them as nonnegative integers, such as 0, 1, 2…. For the GDM, the Q-matrix is also sufficiently general to be polytomous or dichotomous; for the cPDCM the entries are dichotomous. The cPDCM differs from the GDM in that the cPDCM contains all possible interactions such that the item behavior is more flexibly and comprehensively parameterized under the cPDCM.

The cPDCM Measurement Model

The cPDCM does not require dummy variables for the polytomous attributes mastery levels where these polytomous attributes follow the same categorical distribution as mentioned in the previous section. Instead, we consider the attribute latent mastery levels as predictors in the measurement model. For example, we assume Item *i* measures Attribute 1 and 2. Each attribute has 3 levels where 0 represents non-mastery, 1 represents intermediate mastery and 2 represents mastery. The item response probability for Item *i* given examinee *e* with attribute profile $\boldsymbol{\alpha}_e$ is

$$\log\frac{P(X_{ei}=1|\boldsymbol{\alpha}_e)}{P(X_{ei}=0|\boldsymbol{\alpha}_e)}=\lambda_{i,0}+\lambda_{i,1,(1)}\alpha_{e1}+\lambda_{i,1,(2)}\alpha_{e2}+\lambda_{i,2,(1,2)}\alpha_{e1}\alpha_{e2} \tag{3.11}$$

where $\lambda_{i,0}$, $\lambda_{i,1,(1)}$, $\lambda_{i,1,(2)}$ and $\lambda_{i,2,(1,2)}$ are the intercept, the main effect for Attribute 1, the main effect for Attribute 2, and the interaction between Attribute 1 and 2 respectively. The only difference in Equation (3.11) and the LCDM equation (Equation (2.4) in Chapter 2) is the possible attribute values. For the LCDM, $\alpha_{ea} \in \{0,1\}$ and for the cPDCM, $\alpha_{ea} \in \{0,1,2\}$.

As a result of the cPDCM parameterization of the effects for polytomous attributes, the increase of the log-odds of the item response is constrained to be the same between any adjacent mastery levels of an attribute. Table 3.4 shows the possible attribute profiles for examinee *e* and the corresponding parameterizations of the log-odds of the item response.

Like the measurement model of the PDCM, the cPDCM has the monotonic constraints for the main effects and interactions to ensure that an examinee who has a higher mastery level of an attribute will have a higher item response probability than an examinee with lower mastery level of the attribute. After applying the analogous simplification as the previous section, we have

$$\lambda_{i,1,(1)} > 0, \qquad \lambda_{i,1,(2)} > 0,$$

$$\lambda_{i,2,(1,2)} > -\lambda_{i,1,(1)}, \qquad \lambda_{i,2,(1,2)} > -\lambda_{i,1,(2)}$$

Compared to the PDCM which has 9 item parameters for item *i*, the cPDCM is much less complex and only contains 4 item parameters. The cPDCM reduces the number of item parameters by constraining all the PDCM main effects of an attribute across different levels to be the same, $\lambda_{\mathrm{i},1,(1)} = \lambda^{1}_{i,1,(1)} = \lambda^{2}_{i,1,(1)}$, and all the PDCM interactions across different attributes and their mastery levels to be the same, $\lambda_{i,2,(1,2)} = \lambda^{11}_{i,2,(1,2)} = \lambda^{12}_{i,2,(1,2)} = \lambda^{21}_{i,2,(1,2)} = \lambda^{22}_{i,2,(1,2)}$. The general form of the cPDCM is the same as the LCDM except the latent attribute mastery level can be 0, 1, …, $(l_a - 1)$ instead of only 0 and 1. For each item *i*, the number of item parameters needs to be estimated is 2^{A_i}, where A_i is the number of attributes measured by item *i*.

The cPDCM Structural Model

Extending the cPDCM beyond unidimensional models requires specifying a structural model that will incorporate parameters to model the relationship between attributes and attribute levels. We propose the log-linear structural model that uses the same latent attribute variables as in the measurement model of the cPDCM. Using the same example for which we assume the test measures two attributes and each attribute has three mastery levels: non-mastery, intermediate mastery, and mastery, there are again 9 attribute profiles an examinee might be classified into. We again define ν_c as the proportion of examinees being in the attribute profile *c*. To express ν_c on the probability scale as a function of μ_c, we use the same equation as the structural model for the PDCM to describe the relationship, where

$$\nu_c = \frac{\exp(\mu_c)}{\sum_{c'=1}^{9} \exp(\mu_{c'})} \tag{3.12}$$

where μ_c has the formualtion

$$\mu_c = \log\frac{\nu_c}{\nu_9} = \gamma_0 + \gamma_{1,(1)}\alpha_{c1} + \gamma_{1,(2)}\alpha_{c2} + \gamma_{2,(12)}\alpha_{c1}\alpha_{c2} \tag{3.13}$$

α_{c1} and α_{c2} are the mastery levels for Attribute 1 and 2 for attribute profile *c*. The values of α_{c1} and α_{c2} can be 0, 1, 2 representing the non-mastery level, intermediate level and mastery level. γ_0 is the intercept of the structural model, representing the log of the proportion of the non-mastery group and the mastery group; $\gamma_{1,(1)}$ and $\gamma_{1,(2)}$ are the main effects for Attribute 1 and 2, representing the change of the proportion of attribute profile *c* on the log scale compared to the non-mastery group; $\gamma_{2,(12)}$ is the interaction between Attribute 1 and 2 which is present when both attributes in attribute profile *c* are at least intermediate levels.

Since the last attribute profile $\boldsymbol{\alpha_9} = [2\ 2]$ is fixed as the reference group, we have

$$\mu_9 = \log\frac{\nu_9}{\nu_9} = 0 = \gamma_0 + 2 \times \gamma_{1,(1)} + 2 \times \gamma_{1,(2)} + 4 \times \gamma_{2,(12)} \tag{3.14}$$

which is

$$\gamma_0 = -\left(2 \times \gamma_{1,(1)} + 2 \times \gamma_{1,(2)} + 4 \times \gamma_{2,(12)}\right) \tag{3.15}$$

To define the general form of the structural model for the cPDCM, suppose a test measures *A* attributes, the general form for the structural model of the cPDCM is:

$$\mu_c = \gamma_0 + \sum_{a=1}^{A} \gamma_{1,(a)}\alpha_{ca} + \sum_{a=1}^{A-1}\sum_{a'=a+1}^{A} \gamma_{2,(aa')}\alpha_{ca}\alpha_{ca'} + \cdots \tag{3.16}$$

where α_{ca} and $\alpha_{ca'}$ are the mastery levels for Attribute a and a'. The values for α_{ca} and $\alpha_{ca'}$ ranges from 0 to $l_a - 1$ and $l_{a'} - 1$, respectively, where l_a and $l_{a'}$ are the numbers of mastery levels for Attribute a and a'. γ_0 is the intercept; $\gamma_{1,(a)}$ is the main effect for Attribute a; and

$\gamma_{2,(aa')}$ is the interaction between Attribute a and a'. The ellipsis represents the sum of the three-way or higher-order interactions.

We again have the constraint for the parameters of the structural model that stems from setting the kernel of the function equal to zero for the reference group, where

$$\gamma_0 = -\left(\sum_{a=1}^{A}\gamma_{1,(a)}(l_a - 1) + \sum_{a=1}^{A-1}\sum_{a'=a+1}^{A}\gamma_{2,(aa')}\,(l_a - 1)(l_{a'} - 1) + \cdots\right) \tag{3.17}$$

The base rate for each attribute at each level is defined the same as Equation (10) in the previous section.

Table 3.1

Example Dummy Code for Attribute with Three Mastery Levels

Mastery Levels	**Dummy Variables**	
($\boldsymbol{\alpha}_{\mathbf{ea}}$)	α_{ea}^1	α_{ea}^2
0	0	0
1	1	0
2	1	1

Table 3.2

Dummy Variables and Thresholds for Different Mastery Levels

Mastery Levels		**Dummy Variables**				**Conditional Log-odds of Correct Response**
$\boldsymbol{\alpha}_{\mathbf{e1}}$	α_{e2}	α_{e1}^1	α_{e1}^2	α_{e2}^1	α_{e2}^2	$\log\frac{P(X_{ei}=1\|\boldsymbol{\alpha}_e)}{P(X_{ei}=0\|\boldsymbol{\alpha}_e)}$
0	0	0	0	0	0	$\lambda_{i,0}$
0	1	0	0	1	0	$\lambda_{i,0}+\lambda_{i,1,(2)}^1$
0	2	0	0	1	1	$\lambda_{i,0}+\lambda_{i,1,(2)}^1+\lambda_{i,1,(2)}^2$
1	0	1	0	0	0	$\lambda_{i,0}+\lambda_{i,1,(1)}^1$
1	1	1	0	1	0	$\lambda_{i,0}+\lambda_{i,1,(1)}^1+\lambda_{i,1,(2)}^1+\lambda_{i,2,(12)}^{11}$
1	2	1	0	1	1	$\lambda_{i,0}+\lambda_{i,1,(1)}^1+\lambda_{i,1,(2)}^1+\lambda_{i,1,(2)}^2+\lambda_{i,2,(12)}^{11}+\lambda_{i,2,(12)}^{12}$
2	0	1	1	0	0	$\lambda_{i,0}+\lambda_{i,1,(1)}^1+\lambda_{i,1,(1)}^2$
2	1	1	1	1	0	$\lambda_{i,0}+\lambda_{i,1,(1)}^1+\lambda_{i,1,(1)}^2+\lambda_{i,1,(2)}^1+\lambda_{i,2,(12)}^{11}+\lambda_{i,2,(12)}^{21}$
2	2	1	1	1	1	$\lambda_{i,0}+\lambda_{i,1,(1)}^1+\lambda_{i,1,(1)}^2+\lambda_{i,1,(2)}^1+\lambda_{i,1,(2)}^2+\lambda_{i,2,(12)}^{11}$ $+\lambda_{i,2,(12)}^{21}+\lambda_{i,2,(12)}^{22}$

Table 3.3

Attribute Profiles and the Corresponding Threshold

Mastery Levels		**Threshold**
α_{e1}	α_{e2}	$\log\frac{P(X_{ei} = 1\mid\boldsymbol{\alpha}_e)}{P(X_{ei} = 0\mid\boldsymbol{\alpha}_e)}$
0	0	$\lambda_{i,0}$
0	1	$\lambda_{i,0} + \lambda_{i,1,(2)}$
0	2	$\lambda_{i,0} + 2 \cdot \lambda_{i,1,(2)}$
1	0	$\lambda_{i,0} + \lambda_{i,1,(1)}$
1	1	$\lambda_{i,0} + \lambda_{i,1,(1)} + \lambda_{i,1,(2)} + \lambda_{i,2,(1,2)}$
1	2	$\lambda_{i,0} + \lambda_{i,1,(1)} + 2 \cdot \lambda_{i,1,(2)} + 2 \cdot \lambda_{i,2,(1,2)}$
2	0	$\lambda_{i,0} + 2 \cdot \lambda_{i,1,(1)}$
2	1	$\lambda_{i,0} + 2 \cdot \lambda_{i,1,(1)} + \lambda_{i,1,(2)} + 2 \cdot \lambda_{i,2,(1,2)}$
2	2	$\lambda_{i,0} + 2 \cdot \lambda_{i,1,(1)} + 2 \cdot \lambda_{i,1,(2)} + 4 \cdot \lambda_{i,2,(1,2)}$

CHAPTER 4

SIMULATION AND EMPERICAL STUDY DESIGN

To investigate the proposed PDCM and cPDCM, this chapter presents the designs of two simulation studies to investigate the performance of the models and the design of an empirical study to illustrate the application of the PDCM and the cPDCM. The simulation studies will investigate, respectively, 1) the efficacy of both models with respect to the accuracy of parameter estimation and 2) the performance of the models under various conditions including conditions where the models are misspecified. The first simulation study purports to provide a general guidance of the test design, the sample size requirement, and the anticipation of the item parameter estimation and attribute classification accuracy when applying the two PDCMs. The second simulation study help test developers to better understand the trade-offs when selecting between the two models in practice. The empirical study shows an example of using the PDCM and the cPDCM to provide polytomous mastery levels to examinees and the conduct item analysis. The results of the three studies are presented in Chapter 5.

Simulation Study 1: Investigation of PDCM Estimation and Examinee Classification

The simulation study was designed to examine the performance of the PDCM under various conditions. It aims to provide insights to researchers and assessment developers about conditions required for the models to provide accurate diagnostic feedback for examinees' polytomous attribute mastery levels. Results will provide insights into the number of examinees and items required to calibrate a PDCM-based assessment and achieve reasonably accurate item parameters and examinee classifications.

Table 4.1 shows the summary of the simulation conditions. In this study, we manipulated 6 factors: number of attributes (2 levels), number of attribute levels (2 levels), base rate (2 levels), test length (2 levels), test complexity (2 levels), and sample size (4 levels). Crossing these levels yielded 128 conditions. The correlations among attribute pairs were fixed to be .70 across all the simulation conditions. Under all conditions, the generating model was the hybrid PDCM; when the attributes only had two levels, the PDCM is equivalent to the LCDM. The estimation models were always fixed to be the generating model. We conducted 50 replications for each condition. The simulation study was conducted using *Mplus* version 7.4 (Muthén & Muthén, 2012).

Number of Attributes

Conditions included assessments measuring 1, 2 or 3 attributes. Though DCMs are more often designed for assessments measuring more than one attribute (e.g., Bradshaw et al., 2014; Choi, 2009; de la Torre, 2011; Henson, et al., 2009; Kunina-Habenicht, Rupp, & Wilhem, 2012; Madison & Bradshaw, 2014; Bradshaw & Madison, 2015; Templin & Hoffman, 2013), we first evaluate the PDCM under the simplest simulation condition where a test measures only one attribute with more than two mastery levels. One application of a one-attribute test was illustrated using four mastery levels (beginning, basic, developing, proficient) in the context of a large-scale state test (Templin & Bradshaw, 2013) where the UDCM was equivalent to the cPDCM. In addition to the one-attribute conditions, we included conditions with either 2 or 3 attributes. DCM applications with dichotomous attributes have typically investigated between 1 attribute (Templin & Bradshaw, 2013) and 18 attributes (Henson & Templin, 2004) with the majority ranging from 3 to 5 attributes (e.g., Choi, 2009; Henson et al., 2009; de la Torre, 2011; Kunina-Habenicht et al., 2012, Templin & Hoffman, 2013). In our simulation study, our investigation focuses on the smaller end of this range because the PDCM and cPDCM are expected to require larger samples and longer assessments than dichotomous attribute DCMs.

Number of Attribute Levels

For all assessments that measure one, two, or three attributes, we simulated assessments that either contained all attributes with two mastery levels and or all attribute with three mastery levels. We defined the two mastery levels for Attribute a as non-mastery ($\alpha_a = 0$) and mastery ($\alpha_a = 1$). Under the dichotomous attribute conditions, examinees were classified into 2 attribute profiles for the one-attribute assessments, 4 attribute profiles for the two-attribute assessments, and 8 attribute profiles for the three-attribute assessments. When attributes had two mastery levels, the LCDM was used as generating and estimation model. Note that the dichotomous-attribute LCDM is the equivalent to the PDCM when all attributes have two mastery levels. The LCDM item parameter estimates and attribute classification accuracies were then used as a baseline to compare the results of the PDCM under the conditions with three mastery levels. We defined the three attribute mastery levels as non-mastery ($\alpha_a = 0$), intermediate mastery ($\alpha_a = 1$), and mastery ($\alpha_a = 2$). The numbers of attribute profiles for the one-, two-, and three-attribute assessments are 3, 9 and 27 respectively.

Test Length

We examined the test lengths under short and long conditions. In the short test condition, each attribute was measured by at least eight items to yield an accurate and reliable classification (Templin and Bradshaw, 2013). More specifically, the short test length was 8 items for the one-attribute tests, 16 items for the two-attribute tests, and 24 items for the three-attribute tests, where each attribute is measured at least by 8 items. For the long test condition, we doubled the number of items in the test, that is, the one-attribute tests had 16 items, the two-attribute tests had 32 items, and the three-attribute tests had 48 items, where each attribute is measured by twice as many items as in the short test condition, i.e., at least 16 items.

Test Complexity

We simulated tests under two test complexities. The first type of tests contained only *simple items*, meaning each item measured only one attribute. The second type of test contained a blend of simple items and *complex items* which measured two or more attributes. Across all conditions, we simulated the first type of tests to measure either one or three attributes and the second type of tests to measure two or three attributes, such that both types of tests were examined under the same number of conditions.

Tests with simple items. For the one-attribute test conditions, the Q-matrix was a 8×1 matrix and each element equals 1, meaning each item measured the attribute, which was $\mathrm{Q} = (1,1,1,1,1,1,1,1)^{\mathrm{t}}$. Table 4.2 shows the item parameters for the test when the attribute had only two mastery levels. In this case, the PDCM was equivalent to the LCDM. We fixed the intercepts to be -1.250, meaning the item response probability was .223 when an examinee was not a master of the attribute. We fixed the main effects to be 2.250, 2.750, 3.000, 3.500, and these items were repeated twice for the test. The corresponding probabilities of answering these items correctly were .731, .818, .852 and .905 when an examinee was a master of the attribute. We duplicated these 8 items for the long test measuring 16 items.

Table 4.3 shows the item parameters for the one-attribute tests measuring three mastery levels using the same Q-matrix that was used under the conditions for the two-level tests. The intercepts again were fixed to -1.250. The generating model for all items is a blend of the PDCM and the cPDCM, termed the *hybrid* PDCM in this dissertation. Items 1 to 4 were simulated to be like the cPDCM, where the main effects for the intermediate mastery levels and the main effects for the mastery levels to be the same. More specifically, the main effects for Items 1 to 4 were generated as 1.000, 1.250, 1.500 and 1.750. The item response probabilities for examinees with the intermediate mastery level ranged from .438 to .622, and the item response probabilities for examinees with the mastery level ranged from .679 to .905.

We fixed the main effect for the intermediate level and main effect for the mastery level to be different for Items 5 to 8. Thus, the generating model for Items 5 to 8 was the PDCM. Items 5 and 6 had larger main effects compared to Items 7 and 8. This means Item 5 and 6 had higher discriminations and more statistical power to classify examinees into the attribute profiles. Item 5 and Item 7 had smaller main effects for the intermediate levels than the main effects for the mastery levels, and Item 6 and Item 8 had larger main effects for the intermediate level than the main effects for the mastery levels. The item response probabilities for the intermediate mastery group ranged from .378 to .731, and the item response probabilities for the mastery group ranged from .777 to .905. We again repeated the 8 items for the 16-item test.

For the three-attribute assessments with simple items, we fixed Item 1 to Item 8 to measure Attribute 1, Item 9 to Item 16 to measure Attribute 2, and Item 17 to Item 24 to measure Attribute 3. Each attribute was measured by eight items, and the item parameters for the eight items measuring each attribute were fixed the same as in the one-attribute assessments. We developed the test with 48 items measuring three attributes by repeating the 24 items twice.

Tests with a blend of simple and complex items. For tests that contained complex items, the item parameters used in the simulation study are illustrated in Table 4.4 and Table 4.5.

Table 4.4 provides the item parameters for tests measuring two attributes with three mastery levels. The tests contained 16 items with 12 items (Item 1 to Item 12) measuring only one attribute and 4 items (Item 13 to Item 16) measuring two attributes. Most items in the tests were designed to be simple items because simple items only contribute to the classification of a specific attribute while complex items contribute to the classification of all required attributes by the item. Including enough simple items for each attribute can guarantee that the test contains enough information to provide accurate classification for all attributes. Again, the generating model for tests that contained complex items was the hybrid PDCM.

For simple items, the intercepts were -1.250 and the main effects for the intermediate mastery level and the main effects for the mastery level ranged from 1.000 to 2.250. The corresponding item response probabilities was .223 for the nonmastery group and ranged from .438 to .731 for the intermediate mastery group and .679 to .905 for the mastery group. The main effects for the intermediate mastery level and mastery level were also designed for all possible situations. More specifically, Item 1 to Item 4 and Item 7 to Item 10 had equal main effects for the intermediate mastery level and mastery level, meaning the corresponding generating model was the cPDCM. Item 5 and Item 11 had smaller main effects for the intermediate mastery levels, and Item 6 and Item 12 had larger main effects for the intermediate mastery levels. Thus, the generating model for Item 5, 6, 11, and 12 was the PDCM.

Item 13 to Item 16 were complex items that measured two attributes. Since these items provided information for both attributes, the item response probabilities for different mastery groups were more spread out. The intercepts were set to be -1.500 with the corresponding item response probabilities for the nonmastery group equal to .182. The main effects for the intermediate mastery level and mastery level were set to be the same for Item 13 to Item 15, meaning the generating model was the cPDCM. The interactions were set to be zero, positive and negative for each item. Item 16 was a complex item that had a larger main effect for the mastery level than for the intermediate mastery level, meaning the generating model was the PDCM.

Table 4.5 shows the item parameters for the assessments that contains complex items measuring three attributes with three mastery levels. The test length was 24 items where Item 1 to Item 18 measured one attribute and Item 19 to Item 24 measured two attributes. Each attribute was measured by 10 items including 6 simple items and 4 complex items. The item parameters were fixed the same as Table 4.3 for each attribute for the simple items. The interactions for the complex items were again set to be zero, positive or negative.

Sample Sizes

The sample sizes for each condition were 1000, 2000, 5000, and 10000. These sample sizes had broader ranges than the sample sizes used in the previous simulation studies for dichotomous DCMs (e.g., Rupp & Templin, 2008; Bradshaw & Templin, 2012; Cui, Gierl, & Chang, 2012; Kunina-Habenicht, Rupp, & Wilhem, 2012; Madison & Bradshaw, 2014; Bradshaw & Madison, 2015) because the PDCM and the cPDCM are more complex models than the DCMs for dichotomous attributes and are expected to require more examinees to yield an accurate estimation.

Base Rates

The base rate of an attribute in the PDCM and cPDCM represents the proportion of examinees in each mastery level. We examined the estimation of the PDCM under two types of base rates: 1) the base rates were equal across mastery levels; 2) the base rates were not equal across mastery levels. More specifically, the first type of base rates means for the conditions where attributes measured two mastery levels, the base rates are 50% for the non-mastery class and 50% for the mastery class for each attribute. For the condition where attributes measured three mastery levels, the base rates were 33% for the non-mastery class, 33% for the intermediate mastery class, and 34% for the mastery class.

For the second type of base rates, the base rates for attributes with two mastery levels were 30% for the non-mastery group and 70% for the mastery group; the base rates for attributes with three mastery levels were 20%, 30% and 50%.

Evaluation Criteria

We evaluated the performance of the estimation of the dichotomous-attribute LCDM, the PDCM with respect to the convergence rates, the item parameter estimates, and the classification accuracy. The convergence rate for each condition was computed by the percentage of the converged replications compared to the overall number of replications. The item parameter estimates were examined using Root Mean Square Error (RMSE) to measure the extent to which the estimated item parameters are different from the generating values. The classification accuracy was evaluated for each attribute respectively and then for the whole attribute profile.

Convergence Rate. Since the PDCM has more item parameters and classifies examinees into more attribute profile groups compared to the LCDM, the model is more complex and therefore more difficult to converge. For example, consider an item that measures two attributes and each attribute has three mastery levels. The PDCM has 9 item parameters including one intercept, four main effects, and four interactions for the item. Examinees with different mastery levels for each attribute have 9 different item response probabilities. These 9 different probabilities range from 0 to 1. When there are not enough examinees in each attribute profile group, there is limited information available to determine the value of the item response probabilities making convergence difficult to achieve. Therefore, the convergence rate for each condition is an important index to inform the sample size and test design requirements when using the PDCM and the cPDCM.

Accuracy of Item Parameter Estimates. The accuracy of the item parameter estimates was evaluated using the root mean square error (RMSE), defined as

$$\mathrm{RMSE}(\hat{\lambda}) = \sqrt{\frac{1}{\mathrm{R}}\sum_{r=1}^{R}\left(\hat{\lambda} - \lambda\right)^2} \tag{4.1}$$

where λ is the true item parameter, $\hat{\lambda}$ is the estimate of λ, and R is the number of converged replications. RMSE measures the Euclidean distance between the item parameter estimate and the true item parameter. RMSE does not provide information about whether an item parameter is over- or under- estimated. Instead, it provides the average of the bias on the Euclidean distance measure across all the replications.

Classification Accuracy. We computed the classification accuracies for each attribute and the attribute profile. The classification accuracy for an attribute is the percentage of the examinees whose estimated mastery levels are the same as their true mastery level. The classification accuracy for the attribute profile is the percentage of the examinees whose estimated mastery levels for all attributes are the same as their true mastery levels.

Attribute Reliability. Attribute reliability (Bradshaw and Templin, 2013) is a statistic that measures the extent to which an examinee is classified into the same attribute mastery level if he or she theoretically retakes the test. Suppose Attribute a has l_a mastery levels, the steps of computing the attribute reliability for Attribute a is described as follows:

Step 1. Compute the attribute posterior for each examinee, $P(\alpha_{ea} = l|\boldsymbol{X_e})$. The posterior probabilities we obtained from the EM algorithm of M*plus* were used as the probability elements of the categorical distribution.

Step 2. Compute a l_a by l_a table for each examinee where the entries of the table are the corresponding probabilities under the test and re-test classification. Assuming an examinee's test and re-test item responses are independent, we treated the probabilities of being classified into each mastery group for the two tests to be identical. For example, if an examinee's posterior probability of being classified into nonmastery group for Attribute a, $P(\alpha_{ea} = 0|\boldsymbol{X_e})$**,** is equal to p_e^0, the entry (1,1) of the table is $p_e^0 \ \times p_e^0$.

Step 3. Compute the average of the l_a by l_a table entries for all examinees. This table shows the distribution of the attribute posteriors of the population under the theoretical test-retest situation.

Step 4. Compute the polychoric correlation for the l_a by l_a table in Step 3. The polychoric correlation represents how test and retest posteriors are correlated and is considered a measure of how stable the classification is for Attribute a is expected to be across mastery levels.

Simulation Study 2: An Investigation of The Misspecification of Attribute Mastery Levels for The Polytomous-Attribute DCM

The second simulation study is centered on investigating the degree to which: 1) commonly-used model fit indices can detect the over- and under-specification of attribute mastery levels, 2) examinees are misclassified when the mastery levels of an attribute are over-

and under- specified, 3) if examinees were misclassified, the classification of examinees changed compared to their "true" mastery levels.

In this study, we generated the item responses under the LCDM and the hybrid PDCM, as general forms of dichotomous and polytomous DCMs. Specially, we investigated two misspecification conditions: two-level attributes were misspecified as three-level attributes, and three-level attributes were misspecified as two-level attributes. For the first set of conditions, the generating model was the LCDM and the estimation models were the PDCM and the cPDCM. Examinees were generated from two mastery level groups and classified into three mastery levels. Thus, some examinees originally belonging to the mastery or the nonmastery groups might be possibly classified into the intermediate mastery level.

For the second set of conditions, the generating model was the hybrid PDCM and the estimation models were the LCDM and the cPDCM respectively. For the LCDM, examinees were forced to be classified into fewer mastery groups than their generating number of mastery groups. Thus, examinees who were from the intermediate mastery level group were classified into either the nonmastery or mastery group.

Simulation Study Design

The simulation study aimed to investigate the misspecification of the attribute mastery levels. Table 4.6 shows the summary of all conditions. We manipulated the simulation conditions through the following factors: two misspecified attribute mastery levels, two number of attributes measured by the test (2 and 3 attributes), two attribute mastery levels (2 and 3 levels), two test lengths (short and long), four sample sizes (1000, 2000, 5000, 10000), two base rates (equal and unequal), and three estimation models (LCDM, cPDCM, and PDCM). The total number of conditions was 144 and each condition had 50 replications. The simulation study was again conducted by *M*plus version 7.4 (Muthén & Muthén, 2012).

The simulation conditions were similar to Study 1. The main difference lies in that we estimated the simulated responses for each condition under two types of measurement models: LCDM and the hybrid PDCM. The model misspecification was examined when the tests measured one or three attributes with only simple items. The short test length conditions measured each attribute with 8 items, and the longer tests measured each attribute with 16 items. We doubled the short test length items for the long test length conditions. For each test condition, the true item parameters were fixed the same as presented in Study 1. Examinees' base rates were again simulated under the equal and unequal conditions with the values of the proportions the same as in Study 1. The attribute correlations were again fixed to .700.

Generating Model

For the tests that measured attributes with two mastery levels, the item parameters were fixed to be the same as in Table 4.2. Examinees' item responses were generated under the LDCM. For the tests that measured attributes with three mastery levels, the item parameters were fixed to be the same as in Table 4.3. The generating models was the hybrid PDCM. All tests measured either one or three attributes.

Estimation Model

For each simulation condition, the item responses were calibrated by the LCDM with two attribute mastery levels, the PDCM with three mastery levels, and the cPDCM with three mastery levels to investigate the performance of the under-specified and over-specified attribute mastery level conditions. When attributes measured two mastery levels, we simulated the item responses using the LCDM and for estimation models used the LCDM, the over-specified PDCM with three attribute mastery levels, and the overspecified cPDCM also with three attribute mastery levels. Similarly, when attributes measured three mastery levels, we simulated the item responses by the hybrid PDCM with three attribute mastery levels and estimated by the under-specified LCDM, the over-specified PDCM, and the underspecified cPDCM with three attribute mastery levels.

Evaluation Criterion

We evaluated the results by comparing the model convergence rate, model selections under different model fit indices, attribute level classifications, and attribute reliabilities under different model misspecifications.

Model Convergence Rate. We first evaluate the model convergence rates for all model misspecification conditions. Since each condition had 50 replications, the convergence rates were computed as the percentage of converged results among all the replications. This criterion examines whether there is a negative influence on model convergence when the generating model was misspecified.

Model Fit Index. For each condition, we computed three information-based criteria: Akaike's Information Criterion (AIC; Akaike, 1973, 1974, 1987), Bayesian Information Criterion (BIC; Schwarz, 1978) and the sample-size adjusted BIC (SABIC; Sclove, 1987) for all converged replications and summarized the best-fitting model for each index. In our simulation, there were three estimation models including two misspecified models and a "true" model for each generating model. We summarized the model selection results using the percentage that each model was selected under all converged replications. If a replication was not converged, the model selection indices were treated as missing and the model corresponding to the replication was not the best-fitting model. For example, assume AIC was used as the model selection criterion when the generating model was the LCDM and we obtained converged results for a replication when the estimation models was the LCDM or the cPDCM except the PDCM. When summarizing the results, we thus considered the AIC for the PDCM was missing and the PDCM was not the best-fitting model for this replication.

AIC, BIC and SABIC are based on the model likelihood estimation and were adjusted based on the number of model parameters and sample size. The AIC is calculated as

$$\text{AIC} = -2\text{LL} + 2k \tag{4.2}$$

and the BIC is calculated as

$$\text{BIC} = -2\text{LL} + k * \ln(n) \tag{4.3}$$

The SABIC is defined as

$$-2\text{LL} + k * \ln(n^* ((n + 2)/24))$$

where LL represents the loglikelihood, k is the number of model parameters, and n is the sample size.

Attribute Classification. The model misspecification might also influence the attribute classification. We report the average percentages of examinees from each mastery level of the generating model classified into each mastery level under the misspecified model. For example, when the generating model was a blend of the PDCM and the cPDCM, meaning attributes had three mastery levels, the LCDM misspecified attributes as dichotomous and classified examinees into two mastery levels. We compared examinees' true mastery levels and the misspecified mastery levels, especially how examinees who had true intermediate mastery levels were classified into the mastery or nonmastery group in the misspecified condition. Similarly, when the generating model was the LCDM, meaning attributes had two true mastery levels, examinees were classified into three mastery levels and the estimation model was PDCM or cPDCM. We compared the proportions of examinees who originally possessed either nonmastery or mastery level of the two-level attribute were classified into each one of the three mastery levels for the misspecified attributes.

Attribute Reliability. The attribute reliabilities for the LCDM (Bradshaw & Templin, 2013) and for the PDCM introduced in Study 1 were computed respectively when the estimation model classified examinees into two or three mastery levels. The attribute reliabilities measure the extent to which an examinee is expected to be classified into the same attribute mastery level if he or she retakes the test. A higher value of attribute reliability means the classification is more stable.

Empirical Study Design

To demonstrate the application of the PDCM, we analyzed the post-test data from a large-scale mathematics test collected over two years (see also, Madison, 2016; Bottge, Ma, Gassaway, Toland, Butler, & Cho, 2014; Bottge, Toland, Gassaway, Butler, Choo, Griffen, & Ma, 2015). Table 4.7 shows the Q-matrix of the test. The test contained 21 items and measured four mathematics problem-solving skills: ratios and proportional relationships, measurement and data, number system (fractions), and geometry (graphing). Each skill was measured by 4, 6, 5, 6 items respectively. Every item was a simple item where only one attribute was measured by the item. A total of 874 students from Grade 6 to Grade 8 participated in the test.

Measurement Model

In this study, we used the LCDM, PDCM, and cPDCM to analyze students' responses. We estimated the three models using maximum likelihood estimation by *Mplus* 7.4. For all estimation models, we assumed each attribute had either 2 levels: nonmastery and mastery, or 3 levels: nonmastery, intermediate mastery, and mastery. For this four-attribute assessment, there were 16 combinations of mastery levels which was shown in Table 4.8. The estimation models were labeled as LCDM, PDCM-1 to PDCM-15, and cPDCM followed by the mastery levels for each attribute.

The LCDM as an estimation model had all attributes with two mastery levels. Since all items were simple items, each item contained only one intercept and one main effect as item parameters. PDCM-1 to PDCM-15 had at least one attribute with three mastery levels. When an attribute had two mastery levels, the PDCM was equal to the LCDM which contained one intercept and one main effect. When an attribute had three mastery levels, the PDCM had three item parameters per item: one intercept, one main effect for the intermediate mastery level group, and one main effect for the mastery group. We then compared the model fit indices and chose the best fit model among the 16 models. We further compared the best-fitting PDCM with its corresponding cPDCM. Since the main effects are constrained to be equal in the cPDCM, only one intercept and one main effect was estimated for each item.

Evaluation Criteria

Item Characteristic Bar Charts (ICBCs). For the real data analysis, we first evaluated the item parameter estimation accuracy. We illustrate the ICBCs for each item and compared the estimated item response probabilities and the observed proportions of correct responses for each attribute profile. If the estimated item response probabilities and the observed proportion were close to each other, the item parameters for the item might be reasonably estimated and could reflect the behavior of the population answering the item.

Relative Fit. We used Akaike information criterion (AIC; Akaike, 1974), Bayesian information criterion (BIC; Schwarz, 1978), sample-size adjusted BIC, and to evaluate the model fit among the LCDM, PDCM, and cPDCM. Moreover, because the cPDCM was nested in the PDCM, we conducted the likelihood ratio test between the two models to check if the PDCM fitted significantly better than the more parsimonious cPDCM.

Attribute Classification. Finally, we evaluated the classification accuracies for each attribute and each level under the LCDM, PDCM and cPDCM. Because we did not know examinees' "true" mastery levels for each attribute, we only compared the difference of classifications for the three models. For each estimated mastery group, we compared the subscores which were defined as the number of correct items for each attribute. If the classifications and subscores for each mastery level group under the PDCM and cPDCM were similar, the models might yield similar results and it might be reasonable to use the more parsimonious model.

Table 4.1

Summary of Simulation Conditions for Study 1

No. of Attributes	No. of Mastery levels	Test length	Test Complexity	Base Rates	Sample Sizes	Generating Model	Estimation Model	Total No. of Conditions
1	2	8, 16	Simple	Equal, Unequal	1000, 2000, 5000, 10000	LCDM	LCDM	16
	3	8, 16	Simple	Equal, Unequal	1000, 2000, 5000, 10000	PDCM	PDCM	16
2	2	16, 32	Complex	Equal, Unequal	1000, 2000, 5000, 10000	LCDM	LCDM	16
	3	16, 32	Complex	Equal, Unequal	1000, 2000, 5000, 10000	PDCM	PDCM	16
3	2	24, 48	Simple, Complex	Equal, Unequal	1000, 2000, 5000, 10000	LCDM	LCDM	32
	3	24, 48	Simple, Complex	Equal, Unequal	1000, 2000, 5000, 10000	PDCM	PDCM	32
								128

Table 4.2

Item Parameters for One-Attribute Test with Two Mastery Levels

Item	Intercept	Main Effect	Non-mastery IRP	Mastery IRP
1	-1.250	2.250	.223	.731
2	-1.250	2.250	.223	.731
3	-1.250	2.750	.223	.818
4	-1.250	2.750	.223	.818
5	-1.250	3.000	.223	.852
6	-1.250	3.000	.223	.852
7	-1.250	3.500	.223	.905
8	-1.250	3.500	.223	.905

Note. IRP = Item Response Probability.

Table 4.3

Item Parameters for The One-Attribute Test with Three Mastery Levels

Item	Intercept	ME Level 1	ME Level 2	Nonmastery IRP	Intermediate mastery IRP	Mastery IRP	Model
1	-1.250	1.000	1.000	.223	.438	.679	cPDCM
2	-1.250	1.250	1.250	.223	.500	.777	cPDCM
3	-1.250	1.500	1.500	.223	.562	.852	cPDCM
4	-1.250	1.750	1.750	.223	.622	.905	cPDCM
5	-1.250	1.250	2.250	.223	.500	.905	PDCM
6	-1.250	2.250	1.250	.223	.731	.905	PDCM
7	-1.250	0.750	1.750	.223	.378	.777	PDCM
8	-1.250	1.750	0.750	.223	.622	.777	PDCM

Note. IRP = Item Response Probability.

Table 4.4

Item Parameters for the Two-Attribute Complex Tests with Three Attribute Levels

Items		Attribute 1		Attribute 2			Nonmastery IRP	Intermediate Mastery IRP	Mastery IRP	Model
	Intercept	ME Level 1	ME Level 2	ME Level 1	ME Level 2	INX				
1	-1.250	1.000	1.000	0	0	0	.223	.438	.679	cPDCM
2	-1.250	1.250	1.250	0	0	0	.223	.500	.777	cPDCM
3	-1.250	1.500	1.500	0	0	0	.223	.562	.852	cPDCM
4	-1.250	1.750	1.750	0	0	0	.223	.622	.905	cPDCM
5	-1.250	2.000	2.250	0	0	0	.223	.679	.953	PDCM
6	-1.250	2.250	1.250	0	0	0	.223	.731	.905	PDCM
7	-1.250	0	0	1.000	1.000	0	.223	.438	.679	cPDCM
8	-1.250	0	0	1.250	1.250	0	.223	.500	.777	cPDCM
9	-1.250	0	0	1.500	1.500	0	.223	.562	.852	cPDCM
10	-1.250	0	0	1.750	1.750	0	.223	.622	.905	cPDCM
11	-1.250	0	0	1.250	2.250	0	.223	.500	.905	PDCM
12	-1.250	0	0	2.250	1.250	0	.223	.731	.905	PDCM
13	-1.500	1.000	1.000	1.000	1.000	0	.182	-	.924	cPDCM
14	-1.500	.500	.500	.500	.500	.500	.182	-	.924	cPDCM
15	-1.500	1.500	1.500	1.500	1.500	-.500	.182	-	.924	cPDCM
16	-1.500	.500	1.000	.500	1.000	.250	.182	-	.924	PDCM

Note. IRP = Item Response Probability.

Table 4.5

Item Parameters for The Three-Attribute Test with Three Mastery Levels for the Complex Test Design

		Attribute 1		Attribute 2		Attribute 3			Model
Item	Intercept	ME	ME	ME	ME	ME	ME	INX	
		Level 1	Level 2	Level 1	Level 2	Level 1	Level 2		
1	-1.250	1.000	1.000	0	0	0	0	0	cPDCM
2	-1.250	1.250	1.250	0	0	0	0	0	cPDCM
3	-1.250	1.500	1.500	0	0	0	0	0	cPDCM
4	-1.250	1.750	1.750	0	0	0	0	0	cPDCM
5	-1.250	2.000	2.250	0	0	0	0	0	PDCM
6	-1.250	2.250	1.250	0	0	0	0	0	PDCM
7	-1.250	0	0	1.000	1.000	0	0	0	cPDCM
8	-1.250	0	0	1.250	1.250	0	0	0	cPDCM
9	-1.250	0	0	1.500	1.500	0	0	0	cPDCM
10	-1.250	0	0	1.750	1.750	0	0	0	cPDCM
11	-1.250	0	0	2.000	2.250	0	0	0	PDCM
12	-1.250	0	0	2.250	1.250	0	0	0	PDCM
13	-1.250	0	0	0	0	1.000	1.000	0	cPDCM
14	-1.250	0	0	0	0	1.250	1.250	0	cPDCM
15	-1.250	0	0	0	0	1.500	1.500	0	cPDCM
16	-1.250	0	0	0	0	1.750	1.750	0	cPDCM
17	-1.250	0	0	0	0	2.000	2.250	0	PDCM
18	-1.250	0	0	0	0	2.250	1.250	0	PDCM
19	-1.500	1.000	1.000	1.000	1.000	0	0	0	cPDCM
20	-1.500	0.500	0.500	0	0	0.500	0.500	0.500	cPDCM
21	-1.500	0	0	1.500	1.500	1.500	1.500	-0.500	cPDCM
22	-1.500	0.500	1.000	0.500	1.000	0	0	0	PDCM
23	-1.500	0.500	1.000	0	0	0.500	1.000	0.250	PDCM
24	-1.500	0	0	0.500	1.000	0.500	1.000	-0.250	PDCM

Table 4.6

Summary of Simulation Conditions for Study 2

No. of Attributes	No. of Mastery levels	Test length	Test Complexity	Base Rates	Sample Sizes	Generating Model	Estimation Model	Total Number of Conditions
1	2	8, 16	Simple	Equal, Unequal	1000, 2000, 5000, 10000	LCDM	LCDM, PDCM, cPDCM	48
	3	8, 16	Simple	Equal, Unequal	1000, 2000, 5000, 10000	Hybrid PDCM	LCDM, PDCM, cPDCM	48
3	2	24, 48	Simple	Equal, Unequal	1000, 2000, 5000, 10000	LCDM	LCDM, PDCM, cPDCM	48
	3	24, 48	Simple	Equal, Unequal	1000, 2000, 5000, 10000	Hybrid PDCM	LCDM, PDCM, cPDCM	48
								144

Table 4.7

Q-matrix for Empirical Study

ItemNo	Attribute 1	Attribute 2	Attribute 3	Attribute 4
1	1	0	0	0
2	0	1	0	0
3	0	1	0	0
4	0	0	1	0
5	0	0	1	0
6	0	0	1	0
7	0	0	1	0
8	0	0	1	0
9	0	1	0	0
10	0	1	0	0
11	0	1	0	0
12	0	1	0	0
13	1	0	0	0
14	1	0	0	0
15	0	0	0	1
16	0	0	0	1
17	1	0	0	0
18	0	0	0	1
19	0	0	0	1
20	0	0	0	1
21	0	0	0	1
Total	4	6	5	6

Table 4.8

Estimation Models and Attribute Mastery Levels for Empirical Study

Estimation Model	Attribute 1	Attribute 2	Attribute 3	Attribute 4
LCDM	2	2	2	2
PDCM-1	3	2	2	2
PDCM-2	2	3	2	2
PDCM-3	2	2	3	2
PDCM-4	2	2	2	3
PDCM-5	3	3	2	2
PDCM-6	3	2	3	2
PDCM-7	3	2	2	3
PDCM-8	2	3	3	2
PDCM-9	2	3	2	3
PDCM-10	2	2	3	3
PDCM-11	3	3	3	2
PDCM-12	3	3	2	3
PDCM-13	3	2	3	3
PDCM-14	2	3	3	3
PDCM-15	3	3	3	3
cPDCM	3	3	3	3

CHAPTER 5

RESULTS

Results of Simulation Study 1

We present the results for Study 1 under the 128 simulation conditions including: two test complexities, two test lengths (short and long), two base rates (equal and unequal), two numbers of attributes (1 and 2 or 3), two attribute mastery levels (2 and 3), four sample sizes (1000, 2000, 5000, 10000). Results were summarized based on 50 replications for each condition by computing the convergence rate, the item parameter estimation accuracy, attribute classification accuracy, and attribute reliability.

Convergence Rate

Table 5.1 shows the convergence rates for tests that contained only simple items. Note that these tests measured either one or three attributes with two or three mastery levels. The convergence rates for all conditions were quite high and above 98%. There was no meaningful difference across conditions.

When the tests contained both types of simple and complex items, the convergence rates decreased as shown in Table 5.2. The test conditions in Table 5.2 measured 2 or 3 attributes with 2 or 3 mastery levels. In general, the tests that measured two-level attributes had higher convergence rates (ranging from 40% to 100%) than those for the tests that measured three-level attributes (ranging from 0% to 98%). As the sample size and the test length increased, the convergence rates increased for all conditions. When the tests measured two attributes with two mastery levels, the convergence rates maintained a minimum of 96% for all conditions. For other conditions, the convergence rates increased from as low as 0% to above 86% as the sample size increased from 1000 to 10000. More specifically, the convergence rates were 0% when a test measured attributes with three mastery levels by 24 items and the sample size was 1000, and when a test measured attributes with three mastery levels by 48 items across all sample size conditions.

Table 5.3 and 5.4 present the convergence rates when examinees had unequal base rates across attribute mastery levels. The convergence rates for assessments that contain only simple items remained higher than 92% for all conditions. However, the convergence rates shown in Table 5.3 for assessments that contain both simple and complex items were slightly lower compared to Table 5.2 due to having fewer examinees in the nonmastery groups for all attributes. Especially, the convergence rates for the assessments measuring three attributes with three mastery levels were much lower ranging from 0% to 20%. Again, replications did not converge when the assessments contained 48 items and measured three attributes with three mastery levels. This result shows the fewer examinees in some groups brings difficulty to model convergence.

Item Parameter Estimation

Table 5.5 shows the average RMSE for the item parameters across converged replications for assessments measuring one attribute. The average RMSE for assessments where the attribute had two mastery levels ranged from 0.033 to 0.173. Specifically, the RMSE for the main effects, ranging from 0.054 to 0.173, were larger than the RMSEs for the intercepts, ranging from 0.033 to 0.121. Moreover, the RMSE for item parameters of assessments with 16 items, ranging from 0.037 to 0.173, were smaller than those for assessments with 8 items, ranging from 0.033 to

0.072. When the sample size reached 10000, the RMSE for item parameters for both test lengths were more stable around 0.030 for intercepts and 0.050 for main effects. The average RMSE for assessments where the attribute had three mastery levels were larger, ranging from 0.046 to .520, excluding two conditions where RMSE was extremely high as 164.284 and 139.967. In these two conditions, RMSE indicated the model did not converge to a stable value for the main effect for the mastery level. When the test length increased from 8 to 16, the RMSE decreased significantly from larger than 0.500 to smaller than 0.050. Again, the RMSE for the intercepts were the smallest, followed by the main effects for the intermediate mastery level among all the item parameters. The RMSE for the main effects for the mastery level were the largest and most unstable item parameters, where there might exist inflated estimation among replications when the test length was 8, as shown in Table 5.5 for the sample size 1000 and 5000 conditions.

Table 5.6 presents the RMSE for the item parameters for the assessments measuring three attributes where only one attribute was measured by each item. In the two test length conditions, each attribute was measured by 8 and 16 items. The RMSE for assessments measuring two attributes, ranged from 0.022 to 0.179 and shown the similar pattern to those in Table 5.5. On the other hand, the RMSE for assessments measuring attributes with three mastery levels were more stable compared to the assessments measuring only one attributes and ranged from 0.026 to 1.358. Again, the RMSE for assessments measuring three-level attributes were higher than those for assessments measuring two-level attributes. As test length increased from 24 to 48, the RMSE decreased significantly for all simulation conditions.

Table 5.7 illustrates the RMSE for item parameters of assessments that measured two attributes. Since the assessments were a combination of simple items and complex items, the RMSE for each condition was summarized under the two item types. The RMSE for the assessments that measured two attributes were in general smaller than those for the assessments that measured three attributes. As the test length increased from 16 to 32, the RMSE decreased significantly for all conditions. Among all item parameters, main effects for the mastery level had the largest RMSE, followed by interaction terms and main effects for the intermediate mastery level; intercepts yielded the smallest RMSE.

Table 5.8 presents the RMSE for assessments that measured three attributes and contained a mixture of simple items and complex items. Within 120 hours, M*plus* did not yield converged results for all conditions with 48 items and 24 items under sample size of 1000. The RMSE showed the same pattern as those presented in Table 5.7.

Classification Accuracy

Table 5.9 shows the classification accuracy under the one-attribute condition across attribute mastery levels, test lengths and sample sizes. The classification accuracies were .975 and .976 across different sample sizes when the attribute had two mastery levels and the test length was 8. The classification accuracies for two-level attributes increased from .976 to .997 across different sample sizes as the test length increased from 8 to 16. When the attribute measured three mastery levels, the PDCM classified examinees less accurately than when the attribute measured two mastery levels. The classification accuracy slightly increased from .774 to .788 when the test length was 8 and from .891 to .895 when the test length was 16 as the sample size increased from 1000 to 5000. As the test length increased from 8 to 16, the classification accuracy increased more when the attribute had more mastery levels. The results indicate a test length of 8 is long enough to provide accurate classifications for an attribute with dichotomous mastery levels, while the test lengths need to be longer to yield the same level of accuracy as the attribute measured more mastery levels of the attribute.

Table 5.10 shows the classification accuracies for the assessments measuring three attributes with two and three mastery levels and only simple items. The classification accuracies for each three-level attribute were similar to those under the one-attribute, three mastery level condition (Table 5.9). The increase of the sample size did not have a strong influence on the classification accuracies, though classification accuracies increased as the test length increased for both two and three mastery level conditions. The attribute classification accuracies increased from around .980 to .998 for all three attributes when the test length increased from 24 to 48 when attributes had two mastery levels. The attribute classification accuracies when attributes had three mastery levels were lower by about .200 compared to conditions where attributes had two mastery levels, and they increased from around .780 to .900 as the test length increased from 24 to 48.

Table 5.11 and 5.12 illustrates the classification accuracy for the assessments containing complex items. The attribute classification accuracies for the two-attribute assessments with two mastery levels were higher compared to the two-attribute assessments with three mastery levels. For the assessments measuring two attributes with two mastery levels, the attribute classification accuracies were higher than .970 when the test length was 16 and higher than .995 when the test length was 32. For the assessments measuring two attributes with three mastery levels, the attribute classification accuracies increased from around .800 to .900 as the test length increased from 16 to 32.

Table 5.12 shows the attribute classification accuracies for the three attribute assessments with complex items. The attribute classification accuracies for the 24-item and 48-item assessments when attributes had two mastery levels yielded similar results as shown under the same condition in Table 5.11. For the assessments measuring three attributes with three mastery levels, the attribute classification accuracies were around .80 for the 24-item assessments. The 48-item assessments did not provide converge results because of the model complexity.

Reliability

Table 5.13 shows the attribute reliability when assessments measured one attribute across attribute mastery levels, test lengths, and sample sizes. The reliabilities for attributes with two mastery levels were all larger than .990 and higher than those for attributes with three mastery levels, ranging from .863 to .964 for each condition respectively. For assessments measuring attributes with two or three mastery levels, the attribute reliabilities were higher as the test length increased from 8 to 16. More specifically, the reliabilities increased from around .993 to 1.00 when attributes had two mastery levels and from around .870 to .960 when attributes had three mastery levels. The attribute reliabilities maintained similar values across four sample size conditions.

Table 5.14 presents the attribute reliabilities for assessments measuring three attributes. Note that the test lengths for these assessments were 24 or 48 with each attribute measured by 8 or 16 items respectively. The results shown in Table 5.14 were similar to those in Table 5.13 with a slight increase (less than .01) of the attribute reliabilities for each corresponding condition. The attribute reliabilities yielded similar values for all three attributes because each attribute was measured by items with the same design of item parameters.

Table 5.15 and Table 5.16 show the attribute reliability of the two- and three-attribute assessments which contains both simple and complex items. Each attribute was measured by 10 items including 6 simple items and 4 complex items respectively. The attribute reliabilities retained similar values as shown in Table 5.13 and 5.14 where the assessments contained only 8 simple items for each item. The attribute reliabilities in Table 5.15 were not available for the 48-

item assessments that measured three attributes with three mastery levels because those conditions did not have converged estimation due to the model complexity.

Results of Simulation Study 2

Convergence rates

Table 5.17 presents the convergence rates when the assessments measured only one attribute. The base rates for the attribute were equal across mastery levels. Note that the convergence rates were computed based on the results of 50 replications for each condition. For assessments measuring one attribute, the model misspecification conditions included two test lengths, two generating models, two misspecification models, and four sample sizes. Because these assessments measured only one attribute and contained only simple items, the convergence rates were all greater than 96% even when the models were misspecified. Specially, the convergence rates were equal to 100% when the estimation models were the LCDM or the cPDCM. When the LCDM was the generating model and the PDCM was the estimation model, the convergence rates were smaller than 100% under the sample size 5000 and 1000.

Table 5.18 illustrates the model convergence rates for the assessments that measured one attribute with unequal base rates. Specifically, the base rates for the two-level attribute were .30 and .70, and the base rates for the three-level attribute were .20, .30 and .50. The convergence rates were slightly lower than those shown in Table 5.17 and ranged from 94% to 100% across all simulation conditions. Again, the convergence rates were all equal to 100% when the LCDM or cPDCM was used as the estimation model. The PDCM was the most complex model. When the LCDM was misspecified by the PDCM, the convergence rates were 94% for both test lengths conditions with a sample size of 1000.

When the assessments measured more attributes, though contained only simple items, the convergence rates decreased for all simulation conditions, as presented in Table 5.19. The convergence rates were computed within a time limit of 120 hours. If a replication did not yield a converge result within 120 hours, the replication was recorded as nonconverge. When the PDCM was misspecified by the LCDM, meaning all three-level attributes were misspecified as two levels, the model convergence rates were all equal to 100%, except the convergence rate was 98% for the condition with test length 24 and sample size 100. Though the test complexity has increased, the model convergence rates did not significantly decrease when the cPDCM was used as the estimation model, which ranged from 80% to 100%. For conditions where the LCDM was misspecified as the PDCM, the convergence rates were significantly lower compared to other model misspecification conditions, ranging from 10% to 52%.

Table 5.20 illustrates the model convergence rates when the assessments measured three attributes with different base rates for attribute mastery levels. Again, when the LCDM was the generating model, the attribute base rates were .30 for the nonmastery and .70 for the mastery; when the PDCM was the generating model, the attribute base rates were .20 for the nonmastery, .30 for the intermediate mastery and .50 for the mastery. The convergence rates were similar compared to the results shown in Table 5.19. The convergence rates when the cPDCM was used as the estimation model were lower, ranging from 96% to 100%, compared to those when attributes had equal base rates. When the PDCM was the estimation model, the convergence rates ranged from 8% to 92% across all conditions.

Model Selection

In this section, we present the percentage of replications that each model was selected under different simulation conditions when AIC, BIC, and SABIC were used as model selection criterion. The results were computed based on the result of 50 replications. If a replication under a specific model did not yield a converged result, the model selection index was recorded as a missing value and was not used to compare with the same model selection index value under other models. Under each model selection index, we computed the percentages of replications the index preferred each model across the 50 replications under the conditions of two test lengths and four sample sizes.

Table 5.21 presents the model selection results when the LCDM was used as the generating model for assessments that measured only one attribute. The attribute had equal base rates for the nonmastery and mastery group. When AIC was used as the model selection index, the percentages of the “true” model, the LCDM, being selected ranged from 84% to 94%, with those under test length 16 slightly higher (2% to 10%) than those under test length 8. The cPDCM was the second preferable model with the percentage being selected between 6% to 12% and the PDCM was the least preferable model with the percentage being selected between 0% to 6%. BIC and SABIC had higher percentages of accurately selecting the LCDM compared to AIC ranging from 94% to 100%, and lower percentages of inaccurately selecting the cPDCM, ranging from 0% to 6%. The PDCM was not selected as the most favorable model when BIC and SABIC used as the model selection indices. The model selection results were similar as shown in Table 5.22 when the base rates were unequal.

When the hybrid PDCM was the generating model, we considered the LCDM and the cPDCM were the misspecified model, Table 5.23 illustrates the model selection percentages for these three models under the AIC, BIC, and SABIC criteria. AIC preferred to select the PDCM for all simulation conditions with the percentages ranging from 98% to 100%. BIC preferred the LCDM with a percentage equal to 88% when the test length was 8 and sample size was 1000 and had the highest percentages of selecting the PDCM for other conditions ranging from 64% to 100%. SABIC preferred the PDCM for all conditions with percentages ranging from 68% to 100%. As the test length and sample size increased, the percentage of selecting the PDCM also increased for all conditions.

Table 5.24 presents the model selection results when the attribute base rates were unequal. AIC yielded similar results as those when the attribute base rates were equal, with the percentages of selecting the PDCM between 94% and 100%. However, the percentages of replications selecting the PDCM when BIC and SABIC were used increased from 0% to 100%, and from 44% to 100%, respectively as the sample size increased from 1000 to 10000. Specifically, when the test length was 8 and the sample size was 1000 or 2000, the BIC preferred the cPDCM with percentages of 78% and 72%. As the test length increased to 16, the BIC and SABIC selected the PDCM more than 78% of the time.

In the real testing scenario, a test usually measures more than one attribute. Table 5.25 to 5.28 shown the model selection results when a test measured three attributes with two mastery levels or three mastery levels. Table 5.25 presents the model selection percentages under the AIC, BIC, and SABIC when the generating model was the LCDM. For all simulation conditions, these three model selection indices preferred the “true” model – the LDCM with the selection percentage equal to 100% across all conditions. Similarly, when the attribute base rates were unequal, the AIC, BIC, and SABIC had the percentage of favoring the LCDM larger than 98% for all conditions.

Table 5.27 and Table 5.28 shows the model selection percentage when the generating model was the hybrid PDCM. Table 5.27 presents the results when the attribute base rates were equal. The AIC, BIC, and SABIC performed well detecting the attribute mastery levels and selected the cPDCM or the PDCM where the attributes had three mastery levels across all replications and simulation conditions. Among the three model selection indices, the AIC had higher percentage of selecting the “true” model – the PDCM, ranging from 64% to 100% under the test length 24 and ranging from 0% to 100% under the test length 48. The BIC was the least accurate index selecting the PDCM, with 0% of replications selecting the PDCM when (a) the sample size was smaller than 5000 when the test length was 24 and (b) the sample size was smaller 2000 when the test length was 48. SABIC was less strict than BIC yet more strict than AIC in selecting the PDCM, with less than 2% of replications selecting the PDCM when the sample size was smaller than 2000 under the test length 24 and 48.

When attribute base rates were unequal, the population can be considered to be distributed in a more complex manner into mastery classes. Thus, the percentages of selecting the PDCM shown in Table 5.28 were generally higher than the results shown in Table 5.27. The percentages of selecting the PDCM by AIC were higher than 86% for all conditions. The increase of the percentage for BIC and SABIC also ranged from 0% to 98%. Again, as the sample size and test length increased, the percentage of selecting the PDCM also increased.

Classification

Attribute classifications under model misspecification were summarized by the combination of the mastery levels under the generating model and the mastery levels under the estimation model. The results for each simulation condition were average across 50 replications. Table 5.29 shows the classification results for the test conditions that measured one attribute with equal base rates, where 50% examinees were nonmasters and 50% examinees were masters, when the generating model was the LCDM and the estimation model was the PDCM. This table presents the attribute classification under 8 simulation conditions: four sample sizes and two test lengths.

The second column represents the true attribute mastery levels under the LCDM, and the second row represents the attribute mastery levels under the PDCM. When the LCDM was misspecified by the PDCM, examinees who were in the nonmastery level of the LCDM were classified again in the nonmastery level of the PDCM with relatively high classification accuracy, which ranged from 48.1% to 48.8% for the test length equal to 8 and 49.8% to 50.0% for the test length equal to 16. For the other 50% of the examinees in the mastery group under the LCDM, only around 1% were classified into the nonmastery group when the test length was 8 and 0.1% were classified into the nonmastery group when the test length was 16. The majority were classified in the mastery group ranging from 33.2% to 38.5% when the test length was 8 and from 29.6% to 35.6% when the test length was 16. In total, 13.2% to 20.5% of examinees were incorrectly classified across all simulation conditions.

Since examinees who were from the true “mastery” group were less accurately classified into the mastery group compared to examinees from the nonmastery group, the distribution of classifications for mastery groups was further investigated. Figure 5.1 presents the distribution of the classification for the mastery group when the test measured one attribute with sample size 1000 and test length 8. The first figure in Figure 5.1 shows the histogram of the percentages that examinees from the true mastery group who were classified into the nonmastery group; the second figure shows the histogram of the percentages examinees from mastery group were classified into the intermediate mastery group; and the third figure shows the histogram of the percentages examinees from mastery group were classified into the true mastery group. The results show that most examinees from the mastery group were classified into the original “mastery” group and few examinees were classified into the true nonmastery group. The classification for the mastery group was more scattered compared to the nonmastery group because the distribution of the classification shown in Figure 5.1 was more spread out.

When the attribute base rates were unequal with 30% of examinees in the nonmastery group under the LCDM, the classification results shown in Table 5.30 were similar to those in Table 5.29.

Table 5.31 and Table 5.32 illustrate the classification results when the generating model was the LCDM and the estimation model was the cPDCM for equal and unequal base rates. In the two simulation conditions, examinees generated under the LCDM were mainly correctly classified into the nonmastery and mastery groups under the cPDCM with only fewer than 0.1% of examinees classified into the intermediate mastery group across all conditions. When the base rates were equal, there were fewer than 1.3% of examinees who were not classified into their “true” mastery levels under the cPDCM. The classification yielded high accuracies across all simulation conditions and slightly increased by about 1% when the test length increased from 8 to 16. When the attribute base rates were unequal as shown in Table 5.32, the classification accuracy were 0.3% higher for the mastery group compared to those for the nonmastery group when the test length was 8. Again, the classification accuracies retained high values across all simulation conditions.

Table 5.33 and Table 5.34 show the classification results when the PDCM was misspecified by the cPDCM, meaning the attribute mastery levels were specified correctly but some of the item parameters were more constrained under the cPDCM. When the attribute base rates were equal, approximately 33% of examinees were generated for each mastery level. As shown in Table 5.33, the classification accuracy of examinees classified correctly into their “true” mastery groups increased as the sample size and test length increased. For the nonmastery group, the percentages of examinees classified accurately increased from 25.9% to 28.1% when the test length was 8 and increased from 30.0% to 30.8% when the test length was 16. The intermediate mastery group had the lowest classification accuracies among all mastery groups, where the percentages of examinees classified in the intermediate mastery group around 22.0% when the test length was 8 and 27.5% when the test length was 16. The mastery group had the highest classification accuracies with around 30.0% of examinees classified accurately when the test length was 8 and around 31.0% of examinees classified accurately when the test length was 16.

When the attribute base rates were unequal, the percentages of examinees generated were 20% for the nonmastery group, 30% for the intermediate mastery group and 50% for the mastery group. The classification accuracies retained more than 17.0% for the nonmastery group and 43.5% for the mastery group, while those for the intermediate mastery group were relatively lower and were larger than 13.7%. Again, the classification accuracies slightly increased ranging from 0.3% to 6.6% as the sample size and test length increased.

Table 5.35 and Table 5.36 present the attribute classification when the PDCM was misspecified by the LCDM, that is, attributes with three mastery levels were misspecified by two mastery levels. The second column represents the generating attribute mastery levels and the second row represents the estimation mastery levels. Examinees from the nonmastery group and the mastery group were classified into the nonmastery and the mastery group under the LCDM respectively, with relatively high accuracy ranging from 31.9% to 33.4%. As the test length increases from 8 to 16, the accuracy increased around 1%. For examinees who were originally in the intermediate mastery group, the LCDM classified the majority into the mastery group with the percentages ranging from 18.0% to 19.8%. Table 5.36 illustrates the classification under the unequal attribute base rates and yielded similar trend as shown in Table 5.35.

Table 5.37 to Table 5.44 illustrates the classification results for the three attribute assessments when the generating model was misspecified. The classification results were the average of the three attributes. The attribute classifications when the LCDM was misspecified by the PDCM are shown in Table 5.37 and Table 5.38. Different from the assessments that measured one attribute, the three-attribute assessments classified examinees who originally belonged to the nonmastery group into either the nonmastery group or the intermediate mastery group when the test length was 24. For example, when the attribute had equal base rates, the percentages of examinees generated from the nonmastery group under the LCDM ranged from 33.1% to 39.1% for being classified into the nonmastery group under the PDCM, and ranged from 9.2% to 15.6% for being classified into the intermediate mastery group. Similarly, when the attribute had unequal base rates, the percentages of examinees from the nonmastery group under the LCDM were around 45.0% for being classified into the nonmastery group, and were around 24.0% for being classified into the intermediate mastery group. Note that the replications for the condition with test length equal to 24 and sample size equal to 1000 did not yield any converged results. As the test length increased to 48, the PDCM classified the majority of examinees from the nonmastery group under the LCDM into the nonmastery group, with over 39.0% for the equal base rates condition and over 64.7% for the unequal base rates condition. Moreover, examinees from the mastery groups under the LCDM were mostly classified into the mastery groups under the PDCM, with larger than 45.7% for the assessments with equal base rates and larger than 28.4% for the assessments with unequal base rates.

When the LCDM was misspecified as the PDCM, the classification results shown in Table 5.39 and Table 5.40 for the three-attribute assessments followed similar pattern as in Table 5.31 and Table 5.32 where examinees from the nonmastery and mastery groups under the LCDM were again classified into the nonmastery and mastery groups under the PDCM. Because the assessments were more complicated under these conditions, the percentages of examinees being classified into the “true” class were around 5% fewer compared to the one-attribute assessments.

Table 5.41 and Table 5.42 illustrate the classification results when the PDCM was misspecified as the cPDCM. Again, the results followed the same pattern as those for the one-attribute assessments shown in Table 5.34 and Table 5.35. The percentage of being classified in correctly for the nonmastery and mastery groups retained relatively high values, larger than 30.0% across all conditions. However, the percentages of examinees being classified accurately in the intermediate mastery group were around 5% lower than those for the one-attribute assessments.

Table 5.43 and Table 5.44 show the attribute classification when the PDCM was misspecified as the LCDM. Examinees in the nonmastery or mastery group under the PDCM were again mainly classified into the nonmastery or mastery group respectively under the LCDM. Examinees who belonged to the intermediate mastery group under the PDCM were classified partly in either the nonmastery group or the mastery group with the percentages for the mastery group ranging from 2% to 6% higher than those for the nonmastery group.

Reliability

This section summarizes the attribute reliabilities when the generating model was misspecified under different test length and sample size conditions. Each reliability result was averaged across the converged results for the 50 replications for each condition. Table 5.45 shows the reliability for the test that measured one attribute with two mastery levels. The attribute was misspecified by the PDCM or the cPDCM which had three mastery levels. The reliabilities for the PDCM and the cPDCM were similar across all simulation conditions. When the test length was 8, the reliabilities were around .910 and increased to around .970 when the test length was 16. The standard deviations of the reliabilities for the PDCM were around .02 lower than those for the cPDCM, meaning the classification of the PDCM was more robust than the cPDCM.

Table 5.46 shows the attribute reliability for the one-attribute assessments when the generating model was the PDCM and was misspecified by the cPDCM or the LCDM. When the estimation model was the LCDM, the reliabilities represent the consistency of the attribute being classified into two mastery levels and were higher than the reliabilities under the cPDCM. The reliabilities ranged from .944 to .947 for test length equal to 8 and increased to between .980 and .991 for test length equal to 16. The reliabilities for the cPDCM were lower because the attribute was classified into three mastery levels, ranging from .827 to .878 for test length equal to 8 and increased to between .936 and .957 for test length equal to 16.

Table 5.47 and Table 5.48 present the reliabilities for the model misspecification when the test measured three attributes. We only present the reliabilities for Attribute 1 because Attribute 2 and 3 were generated under the same simulation conditions and yielded the similar results. Compared to the reliabilities for the one-attribute assessments, the reliabilities for the three-attribute assessments followed the similar pattern with a slight increase for most simulation conditions.

Results for Empirical Study

Relative Fit

Table 5.49 is the comparison of model fit statistics for the 17 models. Among the 16 models with the combinations of two or three attribute mastery levels, the PDCM with all attributes having three mastery levels was the best fitting model in terms of the AIC, BIC, and SABIC indices. Compared to the cPDCM, the PDCM had smaller AIC and SABIC values but a larger the BIC value (.007). Because the constrained PDCM is nested within the PDCM, we further conducted a likelihood ratio test which indicated the PDCM was significantly better than the constrained PDCM ($\chi^2(21) = 142.198, \mathrm{p} < .001$).

Classifications

Figure 5.2 shows classification results for each attribute under the LCDM, PDCM, and cPDCM. The percentages of students being classified into each level for the PDCM and cPDCM were similar across attributes with differences ranging from 0.1% to 13.9%. The intermediate mastery group from the PDCM and cPDCM consisted of 6.5% to 27.5% of students who were in the LCDM non-mastery group and 16.4% to 3.3% of students who were in the LCDM mastery group.

Figure 5.3 shows the mean subscores for students in each attribute mastery level. The difference of mean subscores between the PDCM and the cPDCM ranges from 0.035 to 0.812 indicating the two models yielded similar results. The mean subscores under the LCDM for the students in the non-mastery group were 0.367 to 0.634 higher than those under the PDCM and cPDCM. Moreover, the mean subscores under the LCDM for the mastery group were 0.355 to 0.608 lower than those under the PDCM and cPDCM. These discrepancies show what we expected to happen: Students with higher scores in the non-mastery group or lower scores in the mastery group under the LCDM were classified into the intermediate mastery group under the PDCM and cPDCM.

Item Characteristic Bar Charts (ICBC)

ICBCs in Figure 5.4 show the item response probabilities (vertical axis) by attribute mastery levels (horizontal axis) for all items in the test. The charts show how the model is functioning with the item response probabilities for the intermediate mastery groups being between those for the mastery groups and non-mastery groups and ranging from .124 to .845. The probabilities of correct response were similar for the PDCM and constrained PDCM with the highest value of .202. As expected, these probabilities in the LCDM for the non-mastery groups were .004 to .092 higher and for the mastery group were .017 to .231 lower when compared to the PDCM models.

Discussion

Simulation Study 1

The benefits of using the LCDM is to provide direct diagnoses to students about which areas or skills needs to be improved. In most cases where the LCDM are used to diagnose students' mastery levels of latent attributes, students can only obtain whether they have mastered the attributes and to what extent we are confident to classify them into the dichotomous mastery levels. Researcher may seek to provide students more detailed feedback rather than master or nonmaster.

In this dissertation, we proposed two models to measure polytomous attributes named the PDCM and the cPDCM. The PDCM and the cPDCM model item response probabilities through a linear combination of the latent attribute variable and item parameters, such as an intercept, main effects and interactions. The saturated PDCM has more flexibility in modeling item response probabilities in that it allows different main effects and interactions across the attributes measured by an item and the mastery levels of these attributes, while conversely, the cPDCM constrains the main effects and interactions of an attribute for different levels to be equal. Such flexibility makes the saturated PDCM has more item parameters to be estimated, and thus requires a larger sample size and long test to achieve an accurate estimation.

Study 1 investigated the item parameter estimation and attribute classification accuracy of the PDCM through various simulation conditions including the number of attributes, the number of attribute mastery levels, test lengths, sample sizes and test complexities. Results showed the PDCM had lower convergence rates than the LCDM when a test measured more than two attributes or contained complex items. The simulation condition with a test length 48 and a complex test design did not yield converged results.

Since the PDCM and the cPDCM are more complex models compared to the LCDM, the item parameter estimates when a test measured attributes with three mastery levels were less accurate than those under the LCDM. As the test length and sample size increased, the item parameters were more accurate with smaller standard deviation. Similarly, since the PDCM classified examinees into three mastery levels, the classification accuracies for the PDCM were lower than those for the LCDM. Again, as the test length increased, the classification accuracies also increased. The classification accuracies did not have significant differences under different test complexities. The results of Study 1 indicate the PDCM and the cPDCM require a longer test length and a larger sample size to yield similar parameter estimation and classification accuracies.

Simulation Study 2

When applying DCMs to provide diagnostic feedback for students' understandings of knowledge, one of the major questions that stakeholders might be interested in is to decide which model to use. The choice of the model can further decide which type of diagnostic information examinees can obtain. The PDCM and the cPDCM broaden the selection of DCMs for stakeholders to decide the number of attribute mastery levels. This study investigated how the use of misspecified model can influence attribute classifications and reliabilities when the item responses were generated from different DCMs.

The simulation studies were conducted under different test lengths, sample sizes, attribute base rates and combinations of generation and estimating models using M*plus* 7.4. Results show that the PDCM, as the most complex model, had the lowest convergence rates (less than 50%) when the test measured three attributes across all conditions. Among all the converged replications, we investigated whether the generating model can be correctly identified by using AIC, BIC, and SABIC. When the LCDM was the generating model, BIC preformed the best to identity the LCDM with percentages higher than 98% across all conditions. When the hybrid PDCM was the generating model, the percentages of selecting either the PDCM or the cPDCM were not as high as selecting the LCDM correctly when the LCDM was the generating model especially when the test measured three attributes. Among all three indices, AIC correctly specified the attribute mastery levels as three and chose either the PDCM or the cPDCM as the best-fitting model with the percentage of larger than 98% for the simulation conditions where the sample size was larger than 2000. Specifically, when the sample size was larger than 2000, the percentage of choosing the PDCM larger than 90%. Moreover, as the sample size and test length increased, the percentages of identifying the "true" model or the correct mastery levels also increased. This shows the model selection indices in general perform well in identifying attribute mastery levels with the percentage larger than 84% when the sample size is larger than or equal to 2000.

In this study, we focused on investigating the attribute classification under different generating and estimation model combinations. When the LCDM was the generating model, both the PDCM and the cPDCM classified most examinees into nonmastery and mastery group. Moreover, because of the item parameter constraints of the cPDCM, fewer examinees were classified into the intermediate mastery group compared to the PDCM. When the hybrid PDCM was the generating model, both the PDCM and cPDCM classified over 80% of examinees into the “true” attribute mastery levels. Moreover, the cPDCM classified examinees accurately into the nonmastery and mastery groups. This is because the PDCM sometimes had very low item parameter estimates for the main effect for the partial mastery group while the cPDCM constraint the main effects to be equal across mastery levels. In this case, fewer examinees were classified into the intermediate mastery levels under the cPDCM compared to the PDCM though the cPDCM was less accurate in the item parameter estimation. When the PDCM was the generating model and the LCDM was the estimation model, examinees from the intermediate mastery group under the PDCM were classified into either nonmastery or mastery group depending on the test design. For example, in our simulation, the one-attribute assessments classified more examinees with the intermediate mastery level into the nonmastery group, while the three-attribute assessments classified more examinees into the mastery group. For the attribute reliabilities, the LCDM always had higher values because it measures fewer mastery levels compared to the PDCM and cPDCM. As the test length increased, the attribute classifications were closer to the generated classes even if the model was misspecified and the attribute reliabilities also increased.

This study suggests that the model selection indices can be used as effective indicators for the detection of attribute mastery levels and model specification. If an attribute mastery level is under specified, examinees will be classified into fewer mastery groups and thus obtain more coarse feedback. On the other hand, if a model is over specified, the classification can be similar to their "true" classes when the test length is long and sample size is large enough.

Empirical Study

The result of the empirical study illustrates an added benefit of the PDCM over the LCDM when the model-data fit supports polytomous attributes: It classifies students into an intermediate mastery group. The model fit indices indicate the PDCM with three attribute mastery levels has the best fit for students' item responses compared to the cPDCM with three attribute mastery levels and the LCDM with dichotomous attribute mastery levels. This type of feedback can be used to more meaningfully group students for differentiated instruction and to accurately signal intermediate mastery when the underlying attribute is more appropriately characterized with multiple levels.

Table 5.1

Convergence Rates for Assessments with Only Simple Items Under Equal Base Rates

Levels	Attributes	Test Length	Sample Size			
			1000	2000	5000	10000
2	1	8	100%	100%	100%	98%
		16	100%	100%	100%	100%
	3	8	98%	100%	98%	100%
		16	100%	100%	100%	100%
3	1	24	100%	100%	100%	100%
		48	100%	100%	100%	100%
	3	24	98%	100%	100%	100%
		48	98%	100%	100%	96%

Note. The convergence rates were based on the M*plus* results with 120 hours.

Table 5.2

Convergence Rates for Assessments with Both Simple and Complex Items Under Equal Base Rates

Levels	Attributes	Test Length	Sample Size			
			1000	2000	5000	10000
2	2	16	96%	100%	100%	100%
		32	100%	96%	96%	98%
	3	24	50%	60%	44%	38%
		48	40%	94%	90%	92%
3	2	16	4%	24%	78%	98%
		32	38%	58%	86%	86%
	3	24	0%	16%	70%	88%
		48	0%	0%	0%	0%

Note. The convergence rates were based on the M*plus* results with 120 hours.

Table 5.3

Convergence Rates for Assessments with Only Simple Items Under Unequal Base Rates

Levels	Attributes	Test Length	Sample Size			
			1000	2000	5000	10000
2	1	8	100%	100%	100%	98%
		16	100%	100%	100%	100%
	3	24	100%	100%	100%	100%
		48	100%	100%	100%	100%
3	1	8	98%	100%	100%	100%
		16	100%	98%	100%	100%
	3	24	98%	92%	98%	98%
		48	100%	100%	88%	44%

Note. The convergence rates were based on the M*plus* results with 120 hours.

Table 5.4

Convergence Rates for Assessments with Both Simple and Complex Items Under Unequal Base Rates

Levels	Attributes	Test Length	Sample Size			
			1000	2000	5000	10000
2	2	16	96%	94%	72%	72%
		32	100%	98%	100%	100%
	3	24	70%	56%	44%	40%
		48	62%	66%	88%	86%
3	2	16	4%	34%	90%	100%
		32	54%	76%	72%	80%
	3	24	0%	6%	8%	20%
		48	0%	0%	0%	0%

Note. The convergence rates were based on the M*plus* results with 120 hours.

Table 5.5

RMSE for Item Parameters of One-attribute Assessments

Sample Size	Attribute Level	Test Length	Intercept	ME1	ME2
1000	2	8	.121	.173	-
		16	.049	.072	-
	3	8	.488	.520	164.284
		16	.161	.227	.265
2000	2	8	.088	.122	-
		16	.036	.051	-
	3	8	.180	.270	.309
		16	.109	.156	.180
5000	2	8	.052	.078	-
		16	.021	.030	-
	3	8	.189	.255	139.967
		16	.065	.093	.108
10000	2	8	.037	.054	-
		16	.033	.051	-
	3	8	.067	.111	.115
		16	.046	.065	.076

Note. ME1 = main effect for the intermediate mastery level; ME2 = main effect for the mastery level.

Table 5.6

RMSE for Item Parameters for Three-attribute Assessments with Only Simple Items

Sample Size	Attribute Level	Test Length	Intercept	ME1	ME2
1000	2	24	.116	.179	-
		48	.047	.073	-
	3	24	1.358	1.337	1.347
		48	.054	.075	.067
2000	2	24	.076	.121	-
		48	.032	.052	-
	3	24	.663	.735	.468
		48	.041	.052	.046
5000	2	24	.050	.079	-
		48	.022	.033	-
	3	24	.086	.137	.156
		48	.026	.034	.028
10000	2	24	.036	.053	-
		48	.035	.051	-
	3	24	.082	.133	.147
		48	.043	.064	.068

Note. ME1 = main effect for the intermediate mastery level; ME2 = main effect for the mastery level.

Table 5.7

RMSE for Item Parameters for Two-attribute Assessments that Contained Complex Items

Sample Size	Attribute Level	Test Length	Item Type	Intercept	ME1	ME2	Interaction
1000	2	16	Simple	.111	.173	-	-
			Complex	.167	.291	-	.407
		32	Simple	.108	.168	-	-
			Complex	.128	.230	-	.351
	3	16	Simple	.209	.323	.376	-
			Complex	.163	.432	.398	.731
		32	Simple	.141	.233	.269	-
			Complex	.174	.345	2.174	.552
2000	2	16	Simple	.081	.127	-	-
			Complex	.112	.188	-	.301
		32	Simple	.077	.117	-	-
			Complex	.090	.169	-	.268
	3	16	Simple	.133	.229	.322	-
			Complex	.162	.401	.783	.741
		32	Simple	.099	.156	.186	-
			Complex	.133	.253	.649	.418
5000	2	16	Simple	.054	.078	-	-
			Complex	.083	.127	-	.188
		32	Simple	.047	.073	-	-
			Complex	.059	.100	-	.158
	3	16	Simple	.080	.158	.178	-
			Complex	.105	.248	.820	.493
		32	Simple	.063	.099	.117	-
			Complex	.078	.159	.275	.275
10000	2	16	Simple	.035	.055	-	-
			Complex	.069	.090	-	.128
		32	Simple	.033	.051	-	-
			Complex	.045	.078	-	.117
	3	16	Simple	.056	.100	.118	-
			Complex	.074	.180	.317	.360
		32	Simple	.045	.068	.082	-
			Complex	.055	.112	.199	.187

Note. ME1 = main effect for the intermediate mastery level; ME2 = main effect for the mastery level.

Table 5.8

RMSE for Item Parameters for Three-attribute Assessments that Contained Complex Items

Sample Size	Attribute Level	Test Length	Item Type	Intercept	ME1	ME2	Interaction
1000	2	24	Simple	.122	.187	-	-
			Complex	.154	.376	-	.675
		48	Simple	.110	.168	-	-
			Complex	.146	.354	-	.630
	3	24	Simple	-	-	-	-
			Complex	-	-	-	-
		48	Simple	-	-	-	-
			Complex	-	-	-	-
2000	2	24	Simple	.082	.124	-	-
			Complex	.106	.319	-	.573
		48	Simple	.076	.122	-	-
			Complex	.097	.295	-	.564
	3	24	Simple	.144	.221	.289	-
			Complex	.192	.358	2.997	2.064
		48	Simple	-	-	-	-
			Complex	-	-	-	-
5000	2	24	Simple	.047	.076	-	-
			Complex	.067	.265	-	.497
		48	Simple	.048	.072	-	-
			Complex	.063	.254	-	.513
	3	24	Simple	.080	.135	.169	-
			Complex	.108	.222	.644	.675
		48	Simple	-	-	-	-
			Complex	-	-	-	-
10000	2	24	Simple	.037	.054	-	-
			Complex	.046	.246	-	.481
		48	Simple	.034	.052	-	-
			Complex	.045	.235	-	.482
	3	24	Simple	.060	.100	.122	-
			Complex	.075	.167	.319	.390
		48	Simple	-	-	-	-
			Complex	-	-	-	-

Note. ME1 = main effect for the intermediate mastery level; ME2 = main effect for the mastery level.

Table 5.9

Classification Accuracies for One-attribute Assessments

Levels	Test Length	Sample Size			
		1000	2000	5000	10000
2	8	.976	.976	.975	.976
	16	.997	.997	.997	.997
3	8	.774	.783	.788	.784
	16	.891	.893	.895	.891

Table 5.10

Classification Accuracies for the Three-attribute Assessments with Only Simple Items

Mastery Level	Test Length	Sample Size	Attribute 1	Attribute 2	Attribute 3	Attribute Profile
2	24	1000	.980	.980	.980	.942
		2000	.981	.980	.981	.944
		5000	.981	.981	.981	.945
		10000	.981	.981	.981	.945
	48	1000	.998	.998	.998	.994
		2000	.998	.998	.998	.993
		5000	.998	.998	.998	.994
		10000	.998	.998	.998	.993
3	24	1000	.774	.784	.780	.504
		2000	.761	.761	.765	.498
		5000	.782	.777	.775	.519
		10000	.810	.809	.809	.554
	48	1000	.902	.902	.902	.740
		2000	.894	.894	.891	.733
		5000	.906	.897	.896	.738
		10000	.907	.907	.907	.752

Table 5.11

Classification Accuracies for the Two-attribute Assessments with Complex Items

Mastery Level	Test Length	Sample Size	Attribute 1	Attribute 2	Attribute Profile
2	16	1000	.973	.979	.953
		2000	.974	.978	.953
		5000	.973	.978	.952
		10000	.974	.979	.953
	32	1000	.996	.998	.994
		2000	.997	.998	.995
		5000	.997	.998	.995
		10000	.996	.998	.994
3	16	1000	.801	.804	.653
		2000	.807	.806	.658
		5000	.815	.815	.672
		10000	.819	.817	.678
	32	1000	.902	.896	.810
		2000	.903	.902	.817
		5000	.906	.905	.823
		10000	.908	.906	.825

Table 5.12

Classification Accuracies for the Three-attribute Assessments with Complex Items

Mastery Level	Test Length	Sample Size	Attribute 1	Attribute 2	Attribute 3	Attribute Profile
2	24	1000	.965	.983	.973	.926
		2000	.966	.983	.973	.925
		5000	.968	.982	.973	.927
		10000	.968	.983	.974	.929
	48	1000	.993	.999	.996	.988
		2000	.994	.998	.996	.988
		5000	.994	.998	.996	.988
		10000	.994	.998	.996	.988
3	24	1000	.805	.778	.808	.524
		2000	.796	.796	.804	.525
		5000	.812	.801	.810	.546
		10000	.814	.806	.812	.551
	48	1000	-	-	-	-
		2000	-	-	-	-
		5000	-	-	-	-
		10000	-	-	-	-

Table 5.13

Attribute Reliability of One-attribute Assessments with Simple Items and Equal Base Rates

Levels	Test length	Sample Size			
		1000	2000	5000	10000
2	8	.994	.993	.993	.993
	16	1.000	1.000	1.000	1.000
3	8	.884	.872	.867	.863
	16	.964	.963	.962	.964

Table 5.14

Attribute Reliability of Three-attribute Assessments with Simple Items and Equal Base Rates

Mastery Level	Test Length	Sample Size	Attribute 1	Attribute 2	Attribute 3
2	24	1000	.996	.996	.996
		2000	.996	.996	.996
		5000	.996	.996	.996
		10000	.996	.996	.996
	48	1000	1.000	1.000	1.000
		2000	1.000	1.000	1.000
		5000	1.000	1.000	1.000
		10000	1.000	1.000	1.000
3	24	1000	.905	.907	.904
		2000	.896	.899	.893
		5000	.894	.894	.894
		10000	.892	.889	.889
	48	1000	.974	.974	.970
		2000	.971	.971	.970
		5000	.969	.972	.969
		10000	.971	.971	.971

Table 5.15

Attribute Reliability of Two-Attribute Assessments with a Mixture of Simple and Complex Items and Equal Base Rates

Mastery Level	Test Length	Sample Size	Attribute 1	Attribute 2
2	16	1000	.992	.995
		2000	.992	.995
		5000	.991	.995
		10000	.992	.995
	32	1000	1.000	1.000
		2000	1.000	1.000
		5000	1.000	1.000
		10000	1.000	1.000
3	16	1000	.908	.913
		2000	.904	.901
		5000	.897	.893
		10000	.895	.890
	32	1000	.974	.972
		2000	.972	.971
		5000	.970	.969
		10000	.970	.968

Table 5.16

Attribute Reliability of Three-Attribute Assessments with a Mixture of Simple and Complex Items and Equal Base Rates

Mastery Level	Test Length	Sample Size	Attribute 1	Attribute 2	Attribute 3
2	24	1000	.988	.998	.993
		2000	.987	.997	.993
		5000	.988	.997	.992
		10000	.987	.997	.992
	48	1000	1.000	1.000	1.000
		2000	1.000	1.000	1.000
		5000	1.000	1.000	1.000
		10000	1.000	1.000	1.000
3	24	1000	.908	.900	.904
		2000	.915	.902	.901
		5000	.902	.894	.898
		10000	.901	.893	.895
	48	1000	-	-	-
		2000	-	-	-
		5000	-	-	-
		10000	-	-	-

Table 5.17

Model Convergence Rates for One-attribute Assessments with Equal Attribute Mastery Base Rates

			Sample Size			
Test Length	Generating Model	Estimation Model	1000	2000	5000	10000
8	Hybrid PDCM	LCDM	100%	100%	100%	100%
		cPDCM	100%	100%	100%	100%
	LCDM	PDCM	100%	100%	98%	96%
		cPDCM	100%	100%	100%	100%
16	Hybrid PDCM	LCDM	100%	100%	100%	100%
		cPDCM	100%	100%	100%	100%
	LCDM	PDCM	100%	100%	98%	98%
		cPDCM	100%	100%	100%	100%

Table 5.18

Model Convergence Rates for One-attribute Assessments with Unequal Attribute Mastery Base Rates

			Sample Size			
Test Length	Generating Model	Estimation Model	1000	2000	5000	10000
8	Hybrid PDCM	LCDM	100%	100%	100%	100%
		cPDCM	100%	100%	100%	100%
	LCDM	PDCM	100%	98%	100%	94%
		cPDCM	100%	100%	100%	100%
16	Hybrid PDCM	LCDM	100%	100%	100%	100%
		cPDCM	100%	100%	100%	100%
	LCDM	PDCM	100%	100%	96%	94%
		cPDCM	100%	100%	100%	100%

Table 5.19

Model Convergence rates for Three-attribute Assessments with Equal Attribute Mastery Base Rates

			Sample Size			
Test Length	Generating Model	Estimation Model	1000	2000	5000	10000
24	Hybrid PDCM	LCDM	98%	100%	100%	100%
		cPDCM	100%	100%	100%	100%
	LCDM	PDCM	10%	52%	34%	34%
		cPDCM	100%	100%	100%	80%
48	Hybrid PDCM	LCDM	100%	100%	100%	100%
		cPDCM	100%	100%	80%	94%
	LCDM	PDCM	32%	26%	44%	20%
		cPDCM	100%	100%	100%	100%

Table 5.20

Model Convergence Rates for Three-attribute Assessments with Unequal Attribute Mastery Base Rates

			Sample Size			
Test Length	Generating Model	Estimation Model	1000	2000	5000	10000
24	Hybrid PDCM	LCDM	100%	100%	100%	100%
		cPDCM	100%	100%	100%	100%
	LCDM	PDCM	8%	50%	36%	26%
		cPDCM	96%	100%	100%	100%
48	Hybrid PDCM	LCDM	100%	100%	100%	100%
		cPDCM	100%	100%	76%	80%
	LCDM	PDCM	92%	72%	32%	16%
		cPDCM	100%	100%	100%	100%

Table 5.21

Model Selection Percentage Under Different Indices for One-attribute Assessments Under the LCDM as Generating Model with Equal Base Rates

Test	Sample	AIC			BIC			SABIC		
Length	Size	PDCM	cPDCM	LCDM	PDCM	cPDCM	LCDM	PDCM	cPDCM	LCDM
8	1000	2%	10%	88%	0%	0%	100%	0%	6%	94%
	2000	4%	12%	84%	0%	0%	100%	0%	4%	96%
	5000	0%	12%	88%	0%	2%	98%	0%	2%	98%
	10000	6%	10%	84%	0%	0%	100%	0%	0%	100%
16	1000	0%	8%	92%	0%	0%	100%	0%	4%	96%
	2000	0%	12%	88%	0%	2%	98%	0%	4%	96%
	5000	2%	12%	86%	0%	0%	100%	0%	2%	98%
	10000	0%	6%	94%	0%	0%	100%	0%	0%	100%

Table 5.22

Model Selection Percentage Under Different Indices for One-attribute Assessments Under the LCDM as the Generating Model with Unequal Base Rates

Test Length	Sample Size	AIC			BIC			SABIC		
		PDCM	cPDCM	LCDM	PDCM	cPDCM	LCDM	PDCM	cPDCM	LCDM
8	1000	2%	10%	88%	0%	0%	100%	0%	6%	94%
	2000	2%	8%	90%	0%	0%	100%	0%	4%	96%
	5000	0%	16%	84%	0%	0%	100%	0%	0%	100%
	10000	4%	10%	86%	0%	2%	98%	0%	2%	98%
16	1000	0%	10%	90%	0%	0%	100%	0%	4%	96%
	2000	0%	8%	92%	0%	0%	100%	0%	0%	100%
	5000	0%	6%	94%	0%	0%	100%	0%	2%	98%
	10000	0%	6%	94%	0%	0%	100%	0%	0%	100%

Table 5.23

Model Selection Percentage Under Different Indices for One-attribute Assessments Under the Hybrid PDCM as the Generating Model with Equal Base Rates

Test	Sample	AIC			BIC			SABIC		
Length	Size	LCDM	cPDCM	PDCM	LCDM	cPDCM	PDCM	LCDM	cPDCM	PDCM
8	1000	2%	0%	98%	88%	6%	6%	30%	2%	68%
	2000	0%	0%	100%	34%	2%	64%	0%	0%	100%
	5000	0%	2%	98%	0%	2%	98%	0%	2%	98%
	10000	0%	0%	100%	0%	4%	96%	0%	0%	100%
16	1000	0%	0%	100%	0%	4%	96%	0%	0%	100%
	2000	0%	0%	100%	0%	0%	100%	0%	0%	100%
	5000	0%	0%	100%	0%	0%	100%	0%	0%	100%
	10000	0%	0%	100%	0%	0%	100%	0%	0%	100%

Table 5.24

Model Selection Percentage Under Different Indices for One-attribute Assessments the PDCM as the Generating Model with Equal Base Rates

Test	Sample	AIC			BIC			SABIC		
Length	Size	LCDM	cPDCM	PDCM	LCDM	cPDCM	PDCM	LCDM	cPDCM	PDCM
8	1000	0%	6%	94%	22%	78%	0%	8%	48%	44%
	2000	0%	0%	100%	4%	72%	24%	2%	28%	70%
	5000	0%	0%	100%	0%	0%	100%	0%	0%	100%
	10000	0%	0%	100%	0%	0%	100%	0%	0%	100%
16	1000	0%	0%	100%	0%	22%	78%	0%	0%	100%
	2000	0%	4%	96%	0%	4%	96%	0%	4%	96%
	5000	0%	0%	100%	0%	0%	100%	0%	0%	100%
	10000	0%	0%	100%	0%	0%	100%	0%	0%	100%

Table 5.25

Model Selection Percentage Under Different Indices for Three-attribute Assessments Under the LDCM as the Generating Model with Equal Base Rates

Test	Sample	AIC			BIC			SABIC		
Length	Size	PDCM	cPDCM	LCDM	PDCM	cPDCM	LCDM	PDCM	cPDCM	LCDM
24	1000	0%	0%	100%	0%	0%	100%	0%	0%	100%
	2000	0%	0%	100%	0%	0%	100%	0%	0%	100%
	5000	0%	0%	100%	0%	0%	100%	0%	0%	100%
	10000	0%	0%	100%	0%	0%	100%	0%	0%	100%
48	1000	0%	0%	100%	0%	0%	100%	0%	0%	100%
	2000	0%	0%	100%	0%	0%	100%	0%	0%	100%
	5000	0%	0%	100%	0%	0%	100%	0%	0%	100%
	10000	0%	0%	100%	0%	0%	100%	0%	0%	100%

Table 5.26

Model Selection Percentage Under Different Indices for Three-attribute Assessments the LCDM as the Generating Model with Unequal Base Rates

Test Length	Sample Size	AIC			BIC			SABIC		
		PDCM	cPDCM	LCDM	PDCM	cPDCM	LCDM	PDCM	cPDCM	LCDM
24	1000	0%	0%	100%	0%	0%	100%	0%	0%	100%
	2000	2%	0%	98%	0%	0%	100%	0%	0%	100%
	5000	0%	0%	100%	0%	0%	100%	0%	0%	100%
	10000	0%	0%	100%	0%	0%	100%	0%	0%	100%
48	1000	0%	0%	100%	0%	0%	100%	0%	0%	100%
	2000	0%	0%	100%	0%	0%	100%	0%	0%	100%
	5000	0%	0%	100%	0%	0%	100%	0%	0%	100%
	10000	0%	0%	100%	0%	0%	100%	0%	0%	100%

Table 5.27

Model Selection Percentage Under Different Indices for Three-attribute Assessments Under the Hybrid PDCM as the Generating Model with Equal Base Rates

Test	Sample	AIC			BIC			SABIC		
Length	Size	LCDM	cPDCM	PDCM	LCDM	cPDCM	PDCM	LCDM	cPDCM	PDCM
24	1000	0%	36%	64%	0%	100%	0%	0%	100%	0%
	2000	0%	10%	90%	0%	100%	0%	0%	98%	2%
	5000	0%	8%	92%	0%	100%	0%	0%	8%	92%
	10000	0%	0%	100%	0%	0%	100%	0%	0%	100%
48	1000	0%	100%	0%	0%	100%	0%	0%	100%	0%
	2000	0%	98%	2%	0%	100%	0%	0%	98%	2%
	5000	0%	0%	100%	0%	0%	100%	0%	0%	100%
	10000	0%	4%	96%	0%	4%	96%	0%	4%	96%

Table 5.28

Model Selection Percentage Under Different Indices for Three-attribute Assessments Under the PDCM as the Generating Model with Unequal Base Rates

Test	Sample	AIC			BIC			SABIC		
Length	Size	LCDM	cPDCM	PDCM	LCDM	cPDCM	PDCM	LCDM	cPDCM	PDCM
24	1000	0%	12%	88%	0%	100%	0%	0%	98%	2%
	2000	0%	8%	92%	0%	100%	0%	0%	94%	6%
	5000	0%	2%	98%	0%	80%	20%	0%	2%	98%
	10000	0%	2%	98%	0%	2%	98%	0%	2%	98%
48	1000	0%	4%	96%	0%	100%	0%	0%	22%	78%
	2000	0%	0%	100%	0%	52%	48%	0%	0%	100%
	5000	4%	10%	86%	4%	10%	86%	4%	10%	86%
	10000	0%	0%	100%	0%	0%	100%	0%	0%	100%

Table 5.29

Classification for One-attribute Test with Equal Base Rates Under the LCDM as Generating Model and the PDCM as Estimation Model

Sample Size	True Mastery Levels	Test length = 8			Test length = 16			True Base Rate
		Nonmastery	Intermediate mastery	Mastery	Nonmastery	Intermediate mastery	Mastery	
1000	Nonmastery	.481	.010	.005	.498	.002	.000	.500
	Mastery	.012	.116	.376	.001	.143	.356	.500
2000	Nonmastery	.484	.007	.007	.500	.001	.000	.500
	Mastery	.012	.106	.385	.001	.167	.330	.500
5000	Nonmastery	.487	.006	.007	.499	.001	.001	.500
	Mastery	.013	.156	.332	.001	.202	.296	.500
10000	Nonmastery	.488	.004	.007	.497	.001	.001	.500
	Mastery	.014	.133	.355	.001	.187	.313	.500

Table 5.30

Classification for One-attribute Test with Unequal Base Rates Under the LCDM as Generating Model and the PDCM as Estimation Model

Sample	True Mastery	Test length = 8			Test length = 16			True Base
Size	Levels	Nonmastery	Intermediate mastery	Rate	Nonmastery	Intermediate mastery	Mastery	Rate
1000	Nonmastery	.283	.011	.500	.300	.001	.001	.300
	Mastery	.009	.233	.500	.001	.222	.475	.700
2000	Nonmastery	.284	.008	.500	.301	.001	.001	.300
	Mastery	.009	.190	.500	.001	.278	.418	.700
5000	Nonmastery	.287	.006	.500	.300	.001	.000	.300
	Mastery	.010	.196	.500	.001	.254	.443	.700
10000	Nonmastery	.286	.006	.500	.298	.001	.001	.300
	Mastery	.010	.230	.500	.001	.241	.459	.700

Table 5.31

Classification for One-attribute Test with Equal Base Rates Under the LCDM as Generating Model and the cPDCM as Estimation Model

Sample Size	Mastery Levels	Test length = 8			Test length = 16			True Base Rate
		Nonmastery	Intermediate mastery	Rate	Nonmastery	Intermediate mastery	Mastery	
1000	Nonmastery	.484	.001	.500	.498	.000	.001	.500
	Mastery	.012	.001	.500	.001	.000	.499	.500
2000	Nonmastery	.486	.000	.500	.500	.000	.001	.500
	Mastery	.013	.000	.500	.001	.000	.497	.500
5000	Nonmastery	.488	.000	.500	.499	.000	.001	.500
	Mastery	.013	.000	.500	.001	.000	.498	.500
10000	Nonmastery	.488	.000	.500	.497	.000	.001	.500
	Mastery	.013	.000	.500	.001	.000	.500	.500

Table 5.32

Classification for One-attribute Test with Unequal Base Rates Under the LCDM as Generating Model and the cPDCM as Estimation Model

Sample Size	Mastery Levels	Test length = 8			Test length = 16			True Base Rate
		Nonmastery	Intermediate mastery	Mastery	Nonmastery	Intermediate mastery	Mastery	
1000	Nonmastery	.284	.001	.300	.300	.000	.001	.300
	Mastery	.009	.001	.700	.001	.000	.697	.700
2000	Nonmastery	.286	.000	.300	.301	.000	.001	.300
	Mastery	.010	.000	.700	.001	.000	.696	.700
5000	Nonmastery	.287	.000	.300	.300	.000	.001	.300
	Mastery	.010	.000	.700	.001	.000	.698	.700
10000	Nonmastery	.287	.000	.300	.298	.000	.001	.300
	Mastery	.010	.000	.700	.001	.000	.700	.700

Table 5.33

Classification for One-attribute Test with Equal Base Rates Under the Hybrid PDCM as Generating Model and the cPDCM as Estimation Model

Sample Size	Mastery Levels	Test length = 8			Test length = 16			True Base Rate
		Nonmastery	Intermediate mastery	Mastery	Nonmastery	Intermediate mastery	Mastery	
1000	Nonmastery	.259	.069	.003	.300	.036	.000	.330
	Intermediate mastery	.069	.222	.043	.041	.273	.017	.330
	Mastery	.002	.037	.297	.000	.016	.318	.340
2000	Nonmastery	.263	.066	.003	.301	.035	.000	.330
	Intermediate mastery	.066	.224	.043	.040	.275	.016	.330
	Mastery	.001	.034	.299	.000	.015	.318	.340
5000	Nonmastery	.268	.063	.003	.299	.035	.000	.330
	Intermediate mastery	.066	.224	.043	.038	.278	.016	.330
	Mastery	.001	.032	.300	.000	.015	.318	.340
10000	Nonmastery	.281	.051	.001	.308	.025	.000	.330
	Intermediate mastery	.053	.218	.063	.026	.274	.033	.330
	Mastery	.001	.055	.279	.000	.032	.302	.340

Table 5.34

Classification for One-attribute Test with Unequal Base Rates Under the Hybrid PDCM as Generating Model and the cPDCM as Estimation Model

Sample Size	Mastery Levels	Test length = 8			Test length = 16			True Base Rate
		Nonmastery	Intermediate mastery	Mastery	Nonmastery	Intermediate mastery	Mastery	
1000	Nonmastery	.171	.028	.001	.182	.022	.000	.200
	Intermediate mastery	.120	.142	.038	.085	.201	.013	.300
	Mastery	.003	.061	.437	.000	.032	.464	.500
2000	Nonmastery	.176	.024	.001	.179	.024	.000	.200
	Intermediate mastery	.123	.138	.039	.075	.213	.013	.300
	Mastery	.003	.059	.437	.000	.032	.463	.500
5000	Nonmastery	.183	.017	.001	.189	.013	.000	.200
	Intermediate mastery	.127	.137	.037	.086	.203	.012	.300
	Mastery	.002	.060	.435	.000	.034	.462	.500
10000	Nonmastery	.183	.018	.001	.190	.011	.000	.200
	Intermediate mastery	.123	.140	.037	.090	.200	.011	.300
	Mastery	.002	.061	.435	.000	.034	.464	.500

Table 5.35

Classification for One-attribute Test with Equal Base Rates Under the Hybrid PDCM as Generating Model and the LCDM as Estimation Model

Sample Size	Mastery Levels	Test length = 8		Test length = 16		True Base Rate
		Nonmastery	Mastery	Nonmastery	Mastery	
1000	Nonmastery	.319	.011	.334	.002	.330
	Intermediate mastery	.148	.186	.132	.198	.330
	Mastery	.007	.329	.001	.333	.340
2000	Nonmastery	.322	.010	.335	.002	.330
	Intermediate mastery	.152	.182	.133	.197	.330
	Mastery	.007	.327	.001	.332	.340
5000	Nonmastery	.324	.009	.333	.001	.330
	Intermediate mastery	.153	.180	.135	.197	.330
	Mastery	.007	.326	.001	.332	.340
10000	Nonmastery	.323	.009	.331	.001	.330
	Intermediate mastery	.151	.183	.135	.198	.330
	Mastery	.007	.327	.001	.334	.340

Table 5.36

Classification for One-attribute Test with Unequal Base Rates Under the Hybrid PDCM as Generating Model and the LCDM as Estimation Model

Sample Size	Mastery Levels	Test length = 8		Test length = 16		True Base Rate
		Nonmastery	Mastery	Nonmastery	Mastery	
1000	Nonmastery	.195	.004	.203	.000	.200
	Intermediate mastery	.209	.091	.252	.048	.300
	Mastery	.016	.485	.005	.492	.500
2000	Nonmastery	.197	.004	.203	.000	.200
	Intermediate mastery	.207	.092	.254	.047	.300
	Mastery	.015	.484	.005	.491	.500
5000	Nonmastery	.197	.004	.202	.000	.200
	Intermediate mastery	.209	.092	.253	.048	.300
	Mastery	.015	.482	.005	.492	.500
10000	Nonmastery	.198	.004	.201	.000	.200
	Intermediate mastery	.207	.093	.253	.048	.300
	Mastery	.015	.483	.004	.494	.500

Table 5.37

Classification for Three-attribute Test with Equal Base Rates Under the LCDM as Generating Model and the PDCM as Estimation Model

Sample Size	Mastery Levels	Test length = 24			Test length = 48			True Base Rate
		Nonmastery	Intermediate mastery	Mastery	Nonmastery	Intermediate mastery	Mastery	
1000	Nonmastery	.391	.092	.010	.390	.100	.007	.500
	Mastery	.005	.006	.496	.003	.015	.486	.500
2000	Nonmastery	.376	.112	.010	.413	.080	.006	.500
	Mastery	.004	.006	.491	.002	.012	.487	.500
5000	Nonmastery	.336	.150	.010	.428	.066	.003	.500
	Mastery	.004	.006	.493	.002	.036	.465	.500
10000	Nonmastery	.331	.156	.010	.454	.040	.003	.500
	Mastery	.004	.006	.493	.002	.044	.457	.500

Table 5.38

Classification for Three-attribute Test with Unequal Base Rates Under the LCDM as Generating Model and the PDCM as Estimation Model

Sample Size	Mastery Levels	Test length = 24			Test length = 48			True Base Rate
		Nonmastery	Intermediate mastery	Mastery	Nonmastery	Intermediate mastery	Mastery	
1000	Nonmastery	.266	.015	.008	.297	.002	.000	.300
	Mastery	.006	.041	.664	.001	.041	.660	.700
2000	Nonmastery	.243	.044	.010	.296	.002	.001	.300
	Mastery	.005	.010	.689	.001	.046	.655	.700
5000	Nonmastery	.245	.040	.010	.295	.001	.001	.300
	Mastery	.005	.018	.682	.001	.075	.627	.700
10000	Nonmastery	.243	.042	.010	.295	.001	.001	.300
	Mastery	.005	.021	.680	.001	.073	.630	.700

Table 5.39

Classification for Three-attribute Test with Equal Base Rates Under the LCDM as Generating Model and the cPDCM as Estimation Model

Sample Size	Mastery Levels	Test length = 24			Test length = 48			True Base Rate
		Nonmastery	Intermediate mastery	Mastery	Nonmastery	Intermediate mastery	Mastery	
1000	Nonmastery	.438	.056	.001	.466	.030	.000	.500
	Mastery	.001	.101	.403	.000	.057	.448	.500
2000	Nonmastery	.441	.055	.000	.471	.029	.000	.500
	Mastery	.001	.100	.402	.000	.054	.446	.500
5000	Nonmastery	.438	.058	.000	.469	.028	.000	.500
	Mastery	.001	.098	.405	.000	.055	.448	.500
10000	Nonmastery	.438	.058	.000	.468	.028	.000	.500
	Mastery	.001	.099	.404	.000	.054	.450	.500

Table 5.40

Classification for Three-attribute Test with Unequal Base Rates Under the LCDM as Generating Model and the cPDCM as Estimation Model

Sample Size	Mastery Levels	Test length = 24			Test length = 48			True Base Rate
		Nonmastery	Intermediate mastery	Mastery	Nonmastery	Intermediate mastery	Mastery	
1000	Nonmastery	.210	.086	.003	.265	.033	.000	.300
	Mastery	.018	.069	.614	.000	.045	.657	.700
2000	Nonmastery	.200	.094	.002	.264	.033	.000	.300
	Mastery	.000	.060	.643	.000	.044	.659	.700
5000	Nonmastery	.216	.078	.001	.262	.033	.000	.300
	Mastery	.000	.066	.638	.000	.043	.663	.700
10000	Nonmastery	.221	.074	.001	.262	.032	.000	.300
	Mastery	.000	.068	.636	.000	.044	.661	.700

Table 5.41

Classification for Three-attribute Test with Equal Base Rates Under the Hybrid PDCM as Generating Model and the cPDCM as Estimation Model

Sample Size	Mastery Levels	Test length = 24			Test length = 48			True Base Rate
		Nonmastery	Intermediate mastery	Mastery	Nonmastery	Intermediate mastery	Mastery	
1000	Nonmastery	.311	.031	.001	.329	.016	.000	.330
	Intermediate mastery	.063	.169	.077	.028	.243	.038	.330
	Mastery	.001	.030	.317	.000	.021	.325	.340
2000	Nonmastery	.316	.031	.001	.330	.016	.000	.330
	Intermediate mastery	.064	.170	.077	.027	.242	.039	.330
	Mastery	.001	.030	.312	.000	.019	.326	.340
5000	Nonmastery	.312	.031	.001	.328	.017	.000	.330
	Intermediate mastery	.060	.173	.078	.028	.243	.039	.330
	Mastery	.001	.029	.316	.000	.019	.326	.340
10000	Nonmastery	.314	.030	.001	.329	.017	.000	.330
	Intermediate mastery	.062	.171	.077	.028	.242	.039	.330
	Mastery	.001	.029	.316	.000	.019	.326	.340

Table 5.42

Classification for Three-attribute Test with Unequal Base Rates Under the Hybrid PDCM as Generating Model and the cPDCM as Estimation Model

Sample Size	Mastery Levels	Test length = 24			Test length = 48			True Base Rate
		Nonmastery	Intermediate mastery	Mastery	Nonmastery	Intermediate mastery	Mastery	
1000	Nonmastery	.203	.023	.001	.211	.013	.000	.200
	Intermediate mastery	.057	.164	.091	.027	.242	.046	.300
	Mastery	.001	.030	.430	.000	.021	.440	.500
2000	Nonmastery	.199	.023	.001	.210	.013	.000	.200
	Intermediate mastery	.054	.166	.094	.025	.244	.045	.300
	Mastery	.001	.029	.434	.000	.021	.442	.500
5000	Nonmastery	.198	.024	.001	.275	.015	.000	.200
	Intermediate mastery	.053	.169	.093	.027	.242	.042	.300
	Mastery	.001	.030	.432	.000	.020	.379	.500
10000	Nonmastery	.198	.024	.001	.228	.013	.000	.200
	Intermediate mastery	.054	.168	.093	.025	.243	.045	.300
	Mastery	.001	.030	.432	.000	.021	.425	.500

Table 5.43

Classification for Three-attribute Test with Equal Base Rates Under the Hybrid PDCM as Generating Model and the LCDM as Estimation Model

Sample Size	Mastery Levels	Test length = 24		Test length = 48		True Base Rate
		Nonmastery	Mastery	Nonmastery	Mastery	
1000	Nonmastery	.336	.007	.343	.002	.330
	Intermediate mastery	.142	.167	.125	.184	.330
	Mastery	.005	.343	.000	.346	.340
2000	Nonmastery	.341	.007	.343	.003	.330
	Intermediate mastery	.140	.170	.120	.188	.330
	Mastery	.005	.337	.000	.345	.340
5000	Nonmastery	.337	.007	.344	.001	.330
	Intermediate mastery	.139	.171	.124	.186	.330
	Mastery	.005	.341	.000	.345	.340
10000	Nonmastery	.337	.007	.345	.001	.330
	Intermediate mastery	.139	.171	.123	.185	.330
	Mastery	.005	.341	.000	.345	.340

Table 5.44

Classification for Three-attribute Test with Unequal Base Rates Under the Hybrid PDCM as Generating Model and the LCDM as Estimation Model

Sample Size	Mastery Levels	Test length = 24		Test length = 48		True Base Rate
		Nonmastery	Mastery	Nonmastery	Mastery	
1000	Nonmastery	.224	.005	.223	.000	.200
	Intermediate mastery	.137	.174	.157	.158	.300
	Mastery	.007	.453	.001	.460	.500
2000	Nonmastery	.218	.005	.222	.000	.200
	Intermediate mastery	.139	.175	.155	.159	.300
	Mastery	.006	.457	.001	.462	.500
5000	Nonmastery	.218	.004	.223	.000	.200
	Intermediate mastery	.141	.174	.156	.158	.300
	Mastery	.007	.457	.001	.462	.500
10000	Nonmastery	.218	.004	.223	.000	.200
	Intermediate mastery	.141	.174	.156	.158	.300
	Mastery	.006	.456	.001	.461	.500

Table 5.45

Reliability for One-attribute Assessments When the LCDM Was Misspecified as the PDCM or the cPDCM

Test length	Sample size	PDCM		cPDCM	
		Equal	Unequal	Equal	Unequal
8	1000	.903(.007)	.912(.006)	.922(.015)	.904(.028)
	2000	.902(.005)	.910(.004)	.920(.020)	.906(.030)
	5000	.901(.003)	.909(.003)	.925(.026)	.905(.032)
	10000	.901(.002)	.909(.002)	.919(.028)	.920(.027)
16	1000	.977(.005)	.977(.002)	.965(.017)	.942(.023)
	2000	.977(.006)	.977(.004)	.967(.017)	.952(.028)
	5000	.977(.006)	.976(.001)	.968(.018)	.948(.034)
	10000	.977(.007)	.976(.001)	.972(.017)	.949(.035)

Table 5.46

Reliability for One-attribute Assessments When the Hybrid PCDM Was Misspecified as the cPDCM or the LCDM

Test length	Sample size	cPDCM		LCDM	
		Equal	Unequal	Equal	Unequal
8	1000	.863(.030)	.839(.036)	.939(.008)	.946(.007)
	2000	.859(.026)	.832(.029)	.937(.006)	.946(.005)
	5000	.859(.022)	.826(.011)	.938(.003)	.947(.003)
	10000	.851(.006)	.827(.008)	.938(.003)	.946(.002)
16	1000	.948(.005)	.940(.018)	.981(.003)	.991(.002)
	2000	.948(.004)	.943(.016)	.981(.002)	.991(.001)
	5000	.947(.002)	.937(.008)	.980(.001)	.991(.001)
	10000	.947(.002)	.936(.005)	.980(.001)	.991(.001)

Table 5.47

Reliability for Three-attribute Assessments When the LCDM Was Misspecified As the PDCM or the cPDCM

Test length	Sample size	PDCM		cPDCM	
		Equal	Unequal	Equal	Unequal
24	1000	.930(.006)	.920(.016)	.922(.005)	.932(.021)
	2000	.924(.024)	.938(.014)	.922(.003)	.933(.021)
	5000	.920(.028)	.924(.017)	.920(.002)	.927(.009)
	10000	.908(.024)	.927(.021)	.920(.001)	.925(.002)
48	1000	.982(.010)	.980(.019)	.972(.003)	.972(.003)
	2000	.985(.006)	.981(.015)	.971(.002)	.972(.002)
	5000	.982(.016)	.972(.022)	.971(.001)	.972(.001)
	10000	.980(.007)	.956(.049)	.971(.001)	.971(.001)

Table 5.48

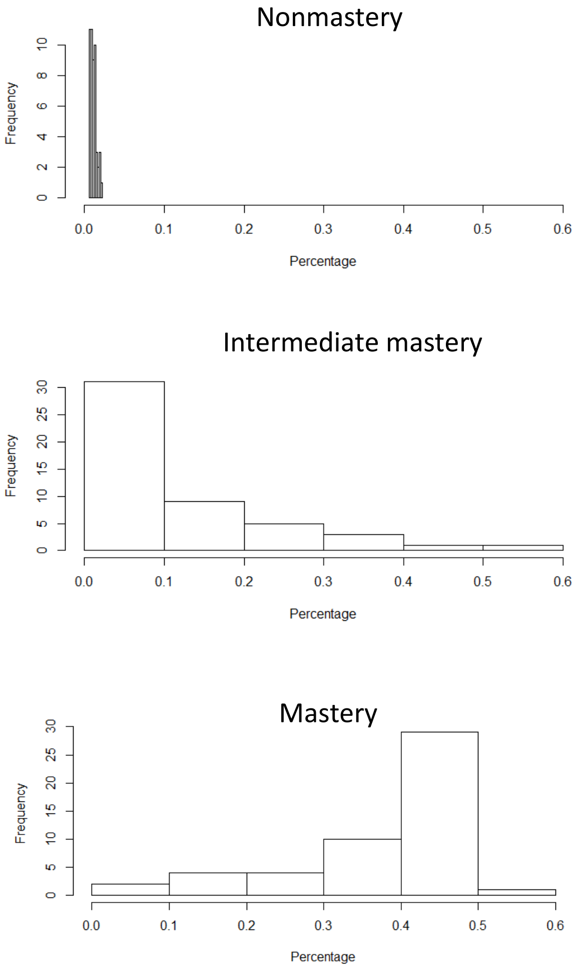
Reliability for Three-attribute Assessments When the Hybrid PCDM Was Misspecified as the cPDCM or the LCDM

Test length	Sample size	cPDCM		LCDM	
		Equal	Unequal	Equal	Unequal
24	1000	.871(.008)	.869(.010)	.961(.006)	.957(.007)
	2000	.870(.006)	.868(.006)	.961(.004)	.955(.005)
	5000	.869(.004)	.867(.005)	.960(.002)	.954(.003)
	10000	.868(.003)	.866(.004)	.960(.002)	.953(.002)
48	1000	.963(.003)	.957(.005)	.986(.002)	.983(.003)
	2000	.961(.001)	.961(.003)	.986(.002)	.982(.002)
	5000	.961(.002)	.962(.002)	.985(.001)	.982(.001)
	10000	.960(.001)	.960(.002)	.985(.001)	.982(.001)

Table 5.49

Summary of Model Fit Statistics for Empirical Study

Model	a1	a2	a3	a4	AIC	BIC	SABIC	LogLikelihood	No. of Parameters
LCDM	2	2	2	2	18731.698	18998.991	18821.148	-9309.849	56
PDCM-1	3	2	2	2	18634.638	18963.981	18744.853	-9248.319	69
PDCM-2	2	3	2	2	18537.569	18886.004	18654.173	-9195.784	73
PDCM-3	2	2	3	2	18569.196	18903.312	18681.008	-9214.598	70
PDCM-4	2	2	2	3	18579.797	18918.686	18693.206	-9218.899	71
PDCM-5	3	3	2	2	NA	NA	NA	NA	NA
PDCM-6	3	2	3	2	18444.731	18855.216	18582.100	-9136.366	86
PDCM-7	3	2	2	3	18455.594	18870.852	18594.560	-9140.797	87
PDCM-8	2	3	3	2	18388.330	18817.907	18532.088	-9104.165	90
PDCM-9	2	3	2	3	18386.182	18815.759	18529.940	-9103.091	90
PDCM-10	2	2	3	3	NA	NA	NA	NA	NA
PDCM-11	3	3	3	2	18256.146	18781.185	18431.851	-9018.073	110
PDCM-12	3	3	2	3	18250.915	18780.726	18428.216	-9014.457	111
PDCM-13	3	2	3	3	18274.879	18799.918	18450.584	-9027.440	110
PDCM-14	2	3	3	3	NA	NA	NA	NA	NA
PDCM-15	3	3	3	3	18087.051	18769.602	18315.467	-8900.526	143
cPDCM	3	3	3	3	18187.249	18769.565	18382.121	-8971.625	122


Nonmastery

Intermediate mastery

Mastery

Figure 5.1 Classification Distribution for True Mastery Group Across All Replications Under One Attribute Condition: Sample Size = 1000, Test Length = 8. Note. Generating model: LCDM, Estimation model: PDCM.

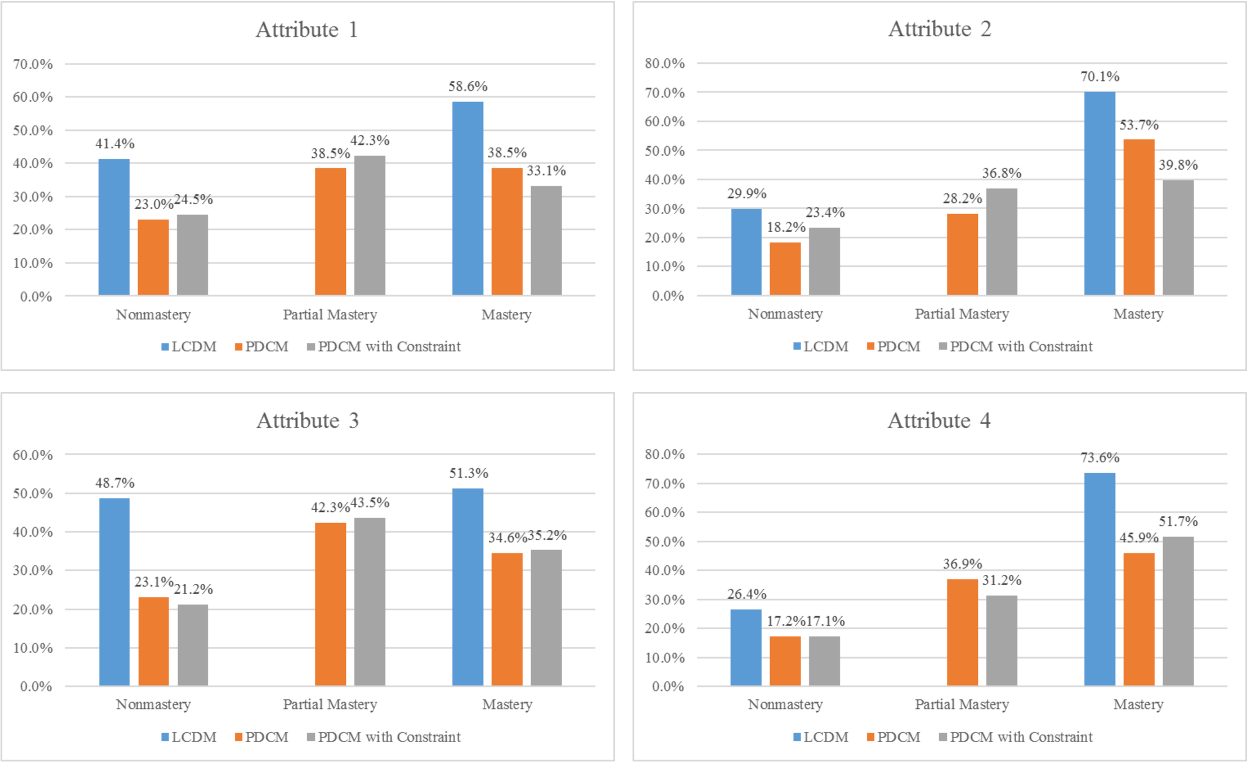


Figure 5.2 Mastery Percentages for Each Attribute under Three Models

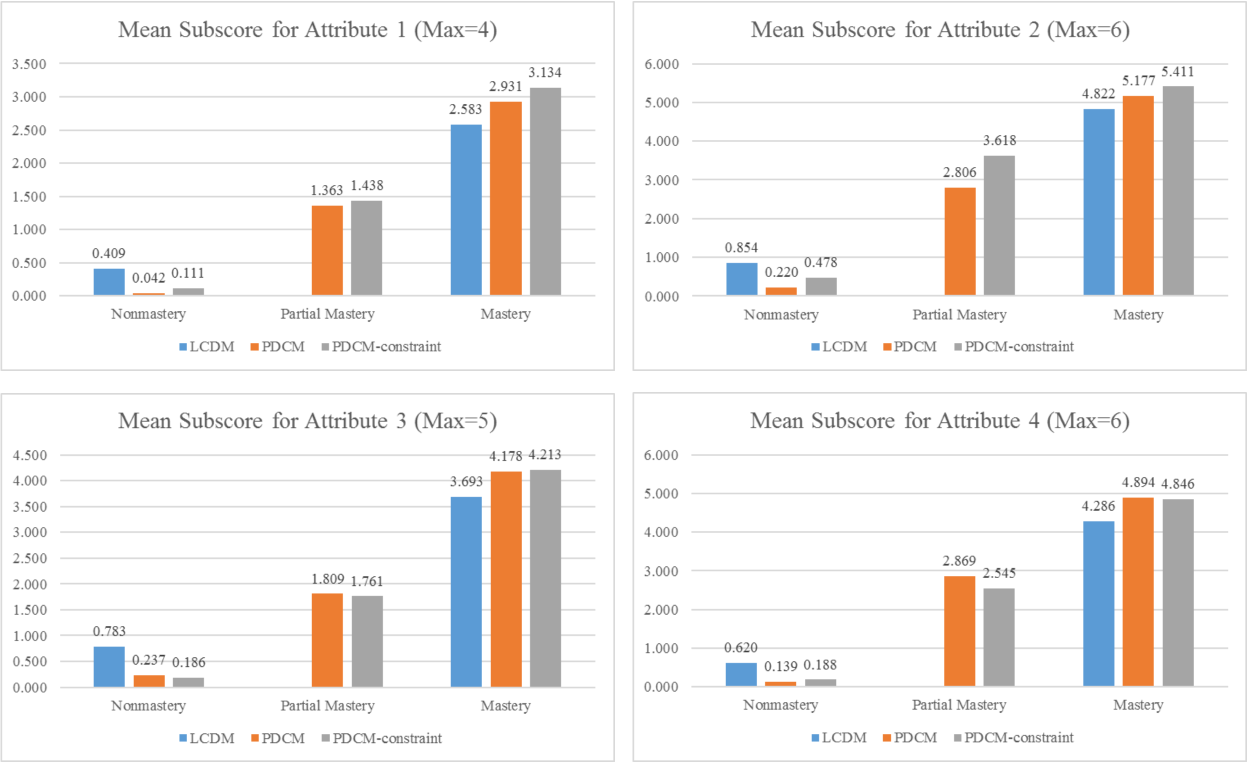
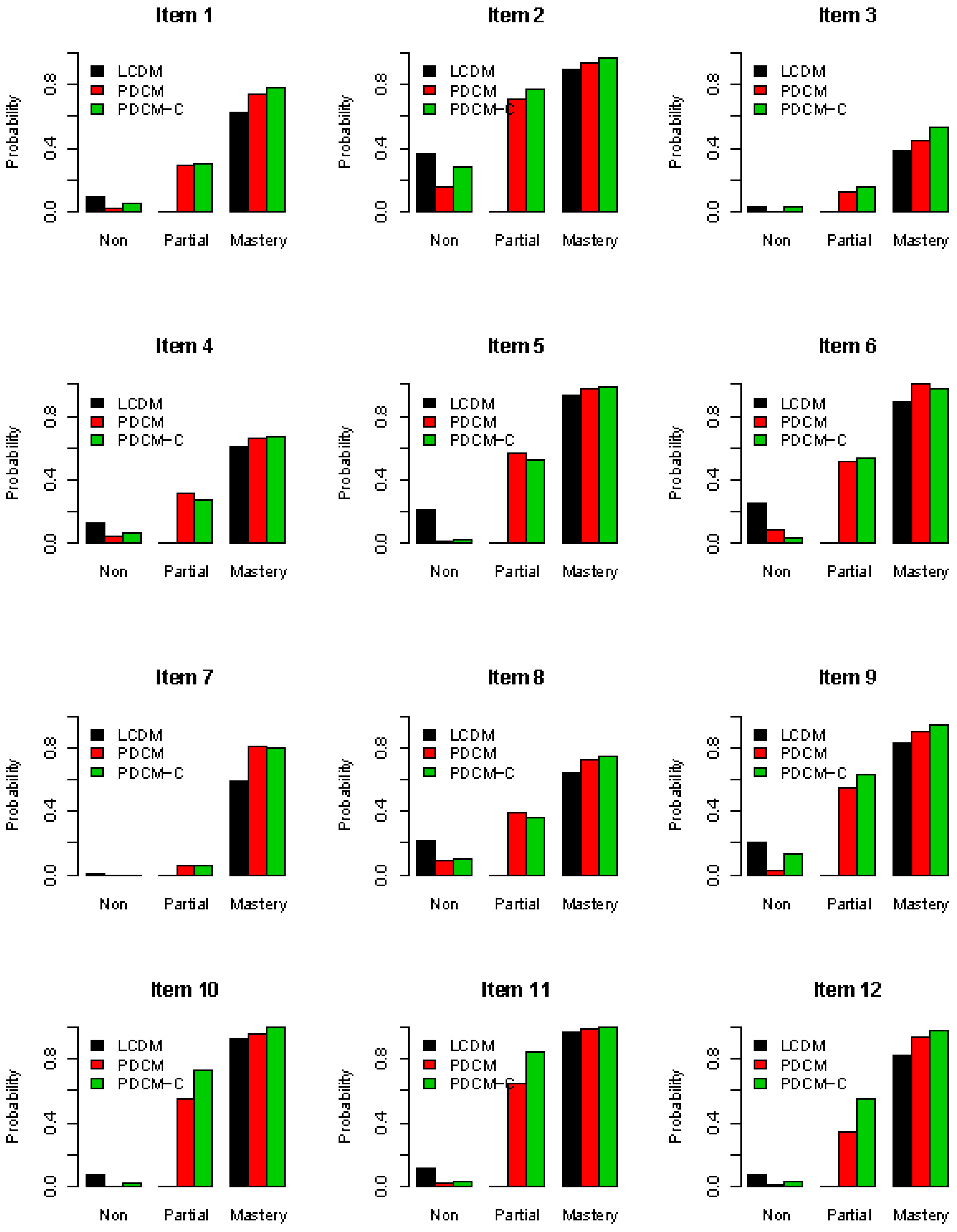


Figure 5.3 Mean Subscore for Each Attribute for Different Mastery Levels


Item 1
Item 2
Item 3
Item 4
Item 5
Item 6
Item 7
Item 8
Item 9
Item 10
Item 11
Item 12
■ LCDM
🟥 PDCM
🟩 PDCM-C
Probability
0.0
0.4
0.8
Non
Partial
Mastery

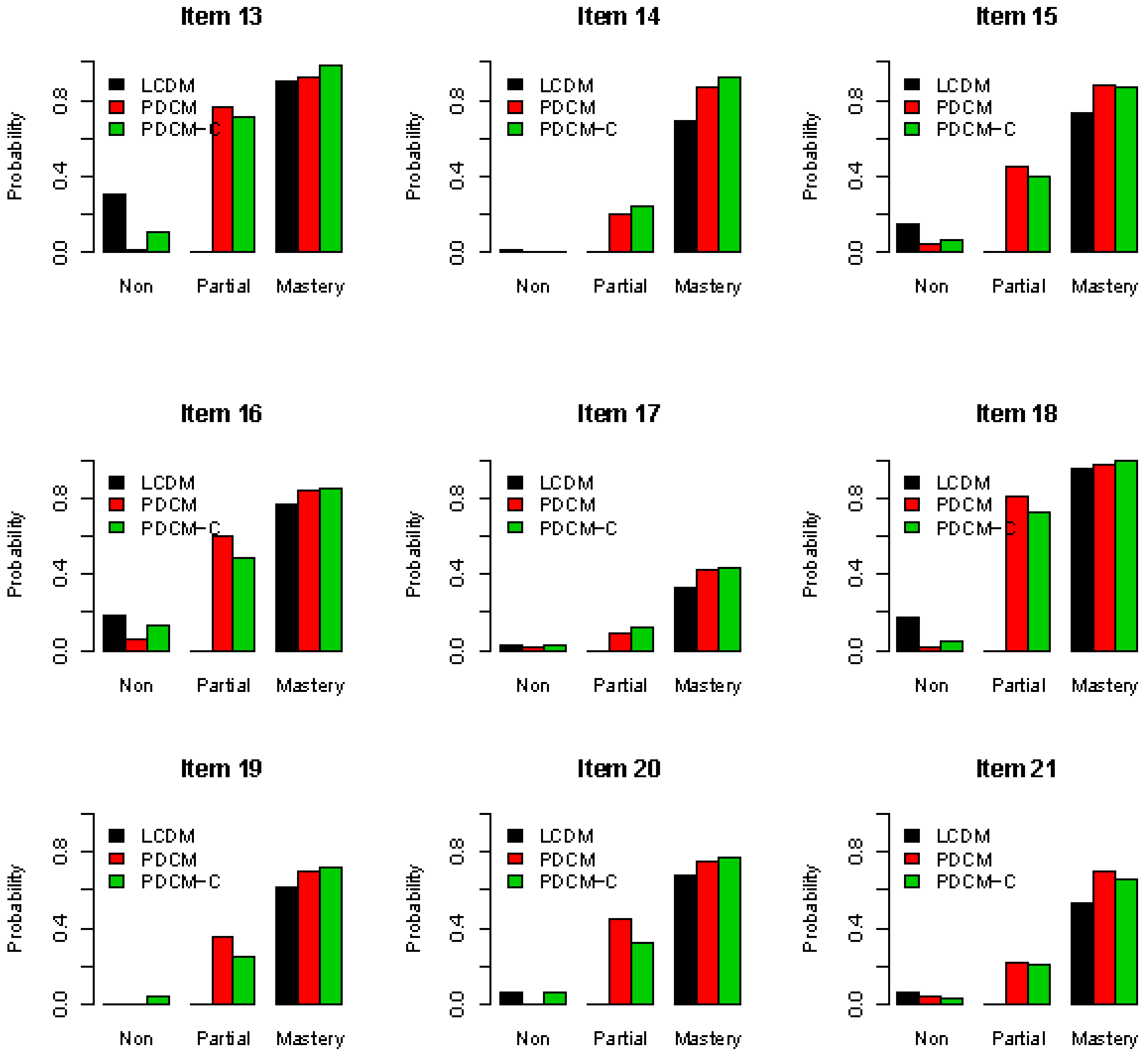


Figure 5.4 Item Characteristic Bar Charts

CHAPTER 6

DISCUSSION

Diagnostic classification models have gained more attention in the application of educational assessments (e.g., Bradshaw et al., 2014; Chiu, Köhn, and Wu, 2016; Kim and Kim, 2013; Kunina-Habenicht, Rupp, and Wilhelm, 2009; Liu et al., 2013; You, et al., 2018) and psychological assessments (e.g., Templin & Henson, 2006) to provide quantitative feedback for multiple attributes. DCMs are a group of item-level probability models that models the probability of answering the item correctly. Many core DCMs have been proposed over past two decades (e.g., Haertel, 1989; Maris, 1999; Junker & Sijtsma, 2001; Hartz, 2002; Templin & Henson, 2006; DiBello, Roussos, & Stout, 2007; de la Torre, 2011). These models can be obtained by constraining the item parameters of the most general models existed in the current literature (von Davier, 2005; Henson et al., 2009; de la Torre, 2010). In this dissertation, the loglinear cognitive diagnostic model (LCDM) was used as the general framework.

One of the limitations of the current DCM literature is that few models can provide feedback to polytomous attributes (Haertel, 1989; Maris, 1999; Junker & Sijtsma, 2001; Hartz, 2002; Templin & Henson, 2006; DiBello, Roussos, & Stout, 2007; de la Torre, 2011). This dissertation focused on addressing this limitation by proposing a new DCM, named the polytomous-attribute DCM (PDCM), that can measure two or more attribute mastery levels. Moreover, the Study 2 in the dissertation investigate the model misspecification on attribute mastery levels.

The Polytomous-attribute Diagnostic Classification Models

There are few studies related to modeling the polytomous-attribute DCMs (e.g., Karalitz, 2004; Templin, 2004; Chen and de la Torre, 2010). These studies mainly focused on developing DCMs under two frameworks: 1) define polytomous attributes at attribute level and specify which level is measured by an item using a Q-matrix (e.g., Karalitz, 2004; Chen and de la Torre, 2010), 2) define polytomous attributes only in order and without specifying a content for each mastery level (e.g., Templin, 2004). The second framework remains the definition of the Q-matrix entries as 0 or 1 as the dichotomous-attribute DCMs where 0 represents an item does not measure an attribute and 1 represents an item measures an attribute. The main differences of the two frameworks are the definition of attribute mastery levels and the design of Q-matrix. Compared to the first framework, the second framework has a more general definition of attribute mastery levels and thus provide a more general attribute diagnostic feedback. This dissertation proposed a general DCM for polytomous attributes under the second framework, termed as the polytomous-attribute DCM (PDCM).

The PDCM was introduced in Chapter 3 and contained two key-components: the measurement model and the structural model. The measurement model is an item-level model for the probability of answering an item correctly under a latent attribute profile. Like the LCDM, the log-odds of the item response function of the PDCM is a linear combination of an intercept, main effects and interactions. The generalizability of the PDCM over the dichotomous DCMs and other polytomous DCMs is that the attribute mastery levels were indicated by a set of latent dummy coded variables and the main effects and interactions were corresponded to each dummy coded variables and combinations. The use of the main effects and interactions for specific attribute mastery levels can capture the relationship between attribute and item response to the most extent. The PDCM also allows the number of mastery levels to vary across attributes. More specifically, if all attributes measured by a test are dichotomous, the PDCM is equal to the LCDM.

The structural model models the attribute space which is the probability of an examinee from the same population being classified into an attribute profile. Among the existing methods, this study generalized the form of loglinear model which is one of the most common structural models of dichotomous DCMs to model the polytomous attribute space. Similarly, the structural model parameters contained intercepts, main effects, and interactions for each dummy variable of attribute mastery levels. Again, the full parameterization of the structural model can capture the attribute mastery of the population to the most extent.

When lack of sample size or test length, the PDCM might provide less accurate item parameter estimation and attribute classification. Chapter 3 also proposed an alternative model to solve this problem, a constrained version of the PDCM named the constrained PDCM (cPDCM). The cPDCM was defined by constraining the main effects and interactions to be equal across attribute mastery levels. Therefore, it contains the same number of item parameters as the LCDM. The difference between the cPDCM and LCDM is that the latent attribute was defined as polytomous values (0, 1, 2…) instead of dichotomous values (0 and 1). Moreover, the cPDCM also has the same forms for the structural model as the LCDM with the difference in defining latent attributes as polytomous values.

Study 1: An Investigation of PDCM Estimation

Two simulation studies and an empirical study were conducted in Chapter 4 and 5. Simulation Study 1 was designed to investigate the model estimation and attribute classification for the PDCM. The purpose of Study 1 is to provide an insight of the test design and the sample size requirements to the PDCM users.Results showed the PDCM functioned properly since it yielded accurate item parameter estimation and high attribute classification accuracy and reliability. In general, a stable and accurate item parameter estimation requires more than 2000 examinees when using the PDCM. As the sample size increased, the RMSEs for the item parameters decreased. An increase of the test length, from 8 items per attribute to 16 items per attribute, can largely decrease the RMSE. Simple items had smaller RMSEs than complex items across all simulation conditions. Compared to the LCDM under the same test length and sample size conditions, the PDCM yielded larger RMSEs than the LCDM due to the more item parameters and more attribute mastery groups.

Attribute classification accuracies for the PDCM were higher than 76% for all simulation conditions. Since the PDCM classified examinees into more attribute mastery classes, the PDCM had lower classification accuracy compared to the LCDM. The increase of the test length can contribute to increasing the classification accuracy under the PDCM. Including complex items in the test can increase the number of items measuring an attribute but did not show significant improvement on the classification accuracy. This shows the simple items may have stronger influence in attribute classification. Attribute reliabilities were higher than .86 for all conditions under the PDCM but smaller then those under the LCDM.

Study 2: An Investigation of Model Misspecification

When applying DCMs to educational or psychological assessments, the test design is required as a priori indicating which attributes are measured by each item. After administering the assessment, the first question for test administrators is to decide which DCM to use for item response calibration. The choice of the model can further decide which type of diagnostic information student might obtain. Study 2 investigated the influence of model misspecification on attribute classifications and reliabilities.

Results from simulation study show that if attribute mastery levels were over specified, most examinees were classified into the same number of the generating classes. This represents the attribute mastery levels can yield similar definitions of knowledge proficiencies. As the test length and sample size increased, more examinees were classified into the major classes. When the attribute mastery levels were under specified, examinees were forced to be classified into fewer classes. The classification under the estimation model depends on the test design and the values of item parameters. In this situation, not only the classification was less stable, but also examinees obtained fewer diagnostic information.

Study 2 also investigated the model selection by three information indices. The model selection indices in general perform well in identifying attribute mastery levels but might need longer test length and larger sample size to identify more complicated models. As the sample size and test length increased, the percentages of identifying the “true” model also increased.

Empirical Study

To demonstrate the application of the PDCM, we analyzed data from a mathematic assessment to diagnose students’ problem-solving skills. We used 17 PDCMs (or LCDM) with different combinations of two or three as attribute mastery levels, where the two mastery levels were defined as nonmastery and mastery, and the three mastery levels were defined as nonmastery, partial mastery and mastery. More specifically, these models contained one LCDM which all attributes had two mastery levels, 15 PDCMs with at least one attribute having three mastery levels, and a cPDCM to compare to the best-fitting PDCM.

The result of the empirical study illustrates an added benefit of the PDCM over the LCDM when the model-data fit supports polytomous attributes: It classifies students into an intermediate mastery group. The model fit indices, AIC, BIC and SABIC indicated the PDCM with all attributes having three mastery levels was the best-fitting model. The attribute classification showed that more than 28% of students were classified into the partial mastery levels under the PDCM for each attribute. The cPDCM yielded similar attribute classifications with less than 14% of difference for attribute mastery levels. When using the LCDM, such information will not be provided because students from the partial mastery group were either classified into the nonmastery group or the mastery group.

The item parameter estimation under the PDCM, cPDCM, and LCDM can further explain the reason for the differences in attribute classification. Among all items, more than 30% items had the differences of the item response probability across attribute mastery levels larger than .30. Most item yielded similar item response probability estimates for the PDCM and cPDCM except around 15% items had differences of the item response probabilities for the partial mastery group larger than .20. Such difference results in the inconsistent of the classification of students. Compared to the PDCM and cPDCM, the LCDM under specified the attribute mastery levels and had higher item response probabilities for the nonmastery group and lower item response probabilities for the mastery group.

Educational Significance

The PDCM contributes the existing DCM literatures as a general DCM for polytomous attributes to provide more detailed diagnostic feedback to educational researchers, teachers and examinees. This type of feedback can be used to more meaningfully group examinees for differentiated instruction and to accurately signal intermediate mastery when the underlying attribute is more appropriately characterized with multiple levels. The flexibility of the PDCM lies in the definition of the attribute mastery levels. It does not require a specific definition for each attribute mastery levels and retains a simple Q-matrix. The PDCM users may decide the number of attribute mastery levels based on the model estimation. It also allows attributes to have different attribute mastery levels so that examinees can obtain the diagnostic feedback of attributes in detail to the different extents. Although using the full PDCM might requires longer test lengths and larger sample sizes, many constraint PDCMs can be obtained to reduce the test length and sample size requirement, such as, by constraining all main effects to be 0, the PDCM becomes a polytomous-attribute DINA model. The submodels of the PDCM provides valuable alternatives for people who seek to obtain more detailed feedback yet struggling with test development or data collection.

Future Study

Though the PDCM and cPDCM show a promising application through the simulation studies and the empirical study, the complexity of these models requires further investigations in many aspects. First, we can further investigate what is the statistically optimum number of mastery levels for each attribute for this data. In Study 1, we only compared the LCDM and PDCMs with three mastery levels for all the attributes. It is possible that some attributes only require to be classified into two mastery levels while the others might prefer to be classified into three or more levels. Second, simulation studies are needed to investigate the accuracy of classification into each mastery level and item parameter estimates, as well as the reliabilities of the attribute classifications under various test conditions. Third, more efforts are needs in providing a guideline for DCM practitioners about how to use the PDCM to provide more detailed feedback to students and educators.

Although many simulation conditions were covered in Study 2, the real testing scenario might be more complicated. One of the limitations of the study is that the assessments contained only simple items, while a real test might contain items that measure two or more attributes. In this situation, the PDCM might become even more complicated when the attribute has three or more mastery levels. Although using the PDCM can provide more detailed feedback to examinees, the model does not always yield converged results. A more constrained model, such as the cPDCM, might be considered as an alternative. Moreover, since the PDCM allows attributes to have different mastery levels, an investigation of a test design that measures multiple attributes with different mastery levels can be conducted as a future study.

REFERENCES

Act, E. S. S. (2015). of 2015. *Public Law*, (114-95), 1177.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Caski (Eds.), *Proceedings of the Second International Symposium on Information Theory* (pp. 267–281). Budapest: Akademiai Kiado.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on* automatic *control*, *19*(6), 716-723.

Akaike, H. (1978). On the likelihood of a time series model.*The Statistician*,**27**, 217–235.

Baker, F. B., & Kim, S. H. (Eds.). (2004). *Item response theory: Parameter estimation techniques*. CRC Press.

Behind, N. C. L. (2001). The elementary and secondary education act. *Public Law print of PL*, 107-110.

Bottge, B. A., Ma, X., Gassaway, L., Toland, M. D., Butler, M., & Cho, S. J. (2014). Effects of blended instructional models on math performance. *Exceptional Children*, *80*(4), 423-437.

Bottge, B. A., Toland, M. D., Gassaway, L., Butler, M., Choo, S., Griffen, A. K., & Ma, X. (2015). Impact of enhanced anchored instruction in inclusive math classrooms. *Exceptional Children*, *81*(2), 158-175.

Bradshaw, L., & Templin, J. (2014, October). *The little model that couldn't: How the DINA model misclassifies students and hides important effects*. Paper presented at the annual meeting of the Northeastern Educational Research Association, Trumbull, CT.

Bradshaw, L., & Templin, J. (2014). Combining item response theory and diagnostic classification models: A psychometric model for scaling ability and diagnosing misconceptions. *Psychometrika*, *79*(3), 403-425.

Bradshaw, L., Izsak, A., Templin, J., & Jacobson, E. (2014). Diagnosing teachers' understandings of rational numbers: Building a multidimensional test within the diagnostic classification framework. *Educational Measurement: Issues and Practice*, 33, 2–14. doi:10.1111/emip.12020

Bradshaw, L. P., & Madison, M. J. (2016). Invariance properties for general diagnostic classification models. *International Journal of Testing*, *16*(2), 99-118.

Briggs, D. C., & Circi, R. (2017). Challenges to the Use of Artificial Neural Networks for Diagnostic Classifications with Student Test Data. *International Journal of Testing*, *17*(4), 302-321.

Buck, G., & Tatsuoka, K.K. (1998). Application of the rule-space procedure to language testing: examining attributes of a free response listening test. *Language Testing*, *15*, 119–157.

Chen, J., & de la Torre, J. (2013). A general cognitive diagnosis model for expert-defined polytomous attributes. *Applied Psychological Measurement*, *37*(6), 419-437.

Chen, Y. H., Ferron, J. M., Thompson, M. S., Gorin, J. S., & Tatsuoka, K. K. (2010). Group comparisons of mathematics performance from a cognitive diagnostic perspective. *Educational Research and Evaluation*, *16*(4), 325-343.

Chiu, C. Y., Köhn, H. F., & Wu, H. M. (2016). Fitting the reduced RUM with Mplus: A tutorial. *International Journal of Testing*, *16*(4), 331-351.

Choi, K. M., Lee, Y. S., & Park, Y. S. (2015). What CDM Can Tell About What Students Have Learned: An Analysis of TIMSS Eighth Grade Mathematics. *Eurasia Journal of Mathematics, Science & Technology Education*, *11*(6).

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart and Winston, 6277 Sea Harbor Drive, Orlando, FL 32887.

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, *69, 333-353.*

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

Darling-Hammond, L., Bae, S., Cook-Harvey, C. M., Lam, L., Mercer, C., Podolsky, A., & Stosich, E. L. (2016). Pathways to new accountability through the Every Student Succeeds Act. *Palo Alto, CA: Learning Policy Institute*.

DiBello, L. V., Roussos, L. A., & Stout, W. (2006). A review of cognitively diagnostic assessment and a summary of psychometric models. *Handbook of statistics*, *26*, 979-1030.

DiBello, L. V., Henson, R. A., & Stout, W. F. (2015). A family of generalized diagnostic classification models for multiple choice option-based scoring. *Applied psychological measurement*, *39*(1), 62-79.

Kim, S. H., & Kim, S. (2013). Incorporating diagnostic aspects to mathematical affects inventory development. *International Journal of Evaluation and Research in Education*, *2*, 163-174.

Korte, G. (2015, December 10th). The Every Student Succeeds Act vs. No Child Left Behind: What's changed? *USA TODAY*. Retrieved from:

https://www.usatoday.com/story/news/politics/2015/12/10/every-student-succeeds-act-vs-no-child-left-behind-whats-changed/77088780/

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2009). A practical illustration of multidimensional diagnostic skills profiling: Comparing results from confirmatory factor analysis and diagnostic classification models. *Studies in Educational Evaluation*, *35*(2-3), 64-70.

Haberman, S. J., von Davier, M., & Lee, Y.-H. (2008). Comparison of multidimensional item response models: Multivariate normal ability distributions versus multivariate polytomous ability distributions (ETS Research Report no. RR-05–16). Princeton, NJ: Educational Testing Service.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* (Vol. 2). Sage.

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

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, *26*(4), 301-321.

Henson, R., & Templin, J. (2005). *Extending cognitive diagnosis models to evaluate the validity of DSM criteria for the diagnosis of pathological gambling.* Poster presented at the National Council for Responsible Gaming: Gambling and Addiction conference, Las Vegas, NV.

Henson, R., & Templin, J. (2007). *Large-scale language assessment using cognitive diagnosis models*. Paper presented at the annual meeting of the National Council for Measurement in Education in Chicago, Illinois.

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.

Im, S., & Yin, Y. (2009). Diagnosing skills of statistical hypothesis testing using the rule space method. *Studies in Educational Evaluation*, *35*(4), 193-199.

Jiang, S., Wang, C., & Weiss, D. J. (2016). Sample size requirements for estimation of item parameters in the multidimensional graded response model. *Frontiers in psychology*, *7*, 109.

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.

Karelitz, T. M. (2004). Ordered category attribute coding framework for cognitive assessments (Unpublished doctoral dissertation). University of Illinois at Urbana–Champaign.

Kim, S. H., & Kim, S. (2013). Incorporating diagnostic aspects to mathematical affects inventory development. *International Journal of Evaluation and Research in Education, 2,* 163–174.

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2009). A practical illustration of multidimensional diagnostic skills profiling: Comparing results from confirmatory factor analysis and diagnostic classification models. Studies in Educational Evaluation, 35, 64–70. doi:10.1016/j.stueduc.2009.10.003

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2012). The Impact of Model Misspecification on Parameter Estimation and Item-Fit Assessment in Log-Linear Diagnostic Classification Models. *Journal of Educational Measurement*, *49*(1), 59-81.

Lee, Y. W., & Sawaki, Y. (2009). Application of three cognitive diagnosis models to ESL reading and listening assessments. *Language Assessment Quarterly*, *6*(3), 239-263.

Lee, Y. S., Park, Y. S., & Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the US national sample using the TIMSS 2007. *International Journal of Testing*, *11*(2), 144-177.

Leighton, J., & Gierl, M. (Eds.). (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge University Press.

Liu, H.-Y., You, X.-F., Wang, W.-Y., Ding, S.-L., & Chang, -H.-H. (2013). The development of computerized adaptive testing with cognitive diagnosis for an English achievement test in China. *Journal of Classification, 30,* 152–172. doi:10.1007/s00357-013-9128-5

Liu, R., Huggins-Manley, A. C., & Bulut, O. (2018). Retrofitting diagnostic classification models to responses from IRT-based assessment forms. *Educational and psychological measurement*, *78*(3), 357-383.

Ma, W., & Torre, J. (2016). A sequential cognitive diagnosis model for polytomous responses. *British Journal of Mathematical and Statistical Psychology*, *69*(3), 253-275.

Madison, M. J., & Bradshaw, L. P. (2015). The effects of Q-matrix design on classification accuracy in the log-linear cognitive diagnosis model. *Educational and Psychological Measurement*, *75*(3), 491-511.

Madison, J. M. (2016). Analyzing pre-test/post-test designs in a diagnostic classification model framework (Unpublished doctoral dissertation). University of Georgia.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, *64*(2), 187-212.

Minchen, N. D., de la Torre, J., & Liu, Y. (2017). A Cognitive Diagnosis Model for Continuous Response. *Journal of Educational and Behavioral Statistics*, *42*(6), 651-677.

Muthén, L. K., & Muthén, B. O. (2012). Mplus statistical modeling software: Release 7.0. *Los Angeles, CA: Muthén & Muthén*.

Reckase, M. D. (2009). Computerized adaptive testing using MIRT. In *Multidimensional Item Response Theory* (pp. 311-339). Springer New York.

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

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

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, *6*(2), 461-464.

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika, 52*, 333–43.

Skaggs, G., Wilkins, J. L., & Hein, S. F. (2016). Grain size and parameter recovery with TIMSS and the general diagnostic model. *International Journal of Testing*, *16*(4), 310-330.

Skaggs, G., Wilkins, J. L., & Hein, S. F. (2017). Estimating an Observed Score Distribution From a Cognitive Diagnostic Model. *Applied psychological measurement*, *41*(2), 150-154.

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. *Diagnostic monitoring of skill and knowledge acquisition*, 453-488.

Templin, J. (2004). Generalized linear proficiency models for cognitive diagnosis (Unpublished doctoral dissertation). University of Illinois at Urbana–Champaign.

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

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

Templin, J. L., Poggio, A., Irwin, P., & Henson, R. (2007, April). *Latent class model based approaches to standard setting.* Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), Chicago, IL.

Templin, J., Rupp, A., Henson, R., Jang, E., & Ahmed, M. (2008). *Nominal response diagnostic models*. Paper presented at the annual meeting of the National Council on Measurement in Education in New York, NY.

Templin, J. L., & Henson, R. A. (2009, April). *Practical issues in using diagnostic estimates: Measuring the reliability of diagnostic estimates.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

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

Templin, J., & Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. Psychometrika, 79(2), 317-339.

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

Von Davier, M. (2005). A general diagnostic model applied to language testing data. *ETS Research Report Series*, *2005*(2).

Wang, C., & Gierl, M. J. (2011). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills in critical reading. *Journal of Educational Measurement*, 48, 165–187. doi:10.1111/jedm.2011.48.issue-2.

Xu, X. & von Davier, M. (2008a). *Fitting the structural general diagnostic model to NEAP data* (RR-08-27). Princeton, NJ: Educational Testing Service.

You, X., Li, M., Zhang, D., & Liu, H. (2018). Application of a Learning Diagnosis System in Chinese Classrooms. *Applied psychological measurement*, *42*(1), 89-94.

Appendix A

M*plus* Code for Tests Measuring Three Attributes With Three Mastery Levels

```
TITLE:  ! Section that appears in header of output file
   DCM for DTMRdata with 4 attributes and full structural model,
28 items, and maximum 2-order item model,
Saturated structural model (Mplus default).

DATA:  ! Location of free format data file
   FILE = l3tl24n1000_1.dat;

VARIABLE:
   NAMES = I1-I24 group;
   USEVARIABLE = I1-I24;
   CATEGORICAL = I1-I24;
   CLASSES = c(27);
   MISSING ARE ALL (99);


ANALYSIS:
   TYPE = MIXTURE;                    ! Estimates latent classes
   STARTS = 0;                   ! Turn off multiple random start feature
   PROCESSORS = 8;                    ! Number of processors available

MODEL:

%OVERALL%

[c#1] (m1) !Latent variable mean for class1
[c#2] (m2) !Latent variable mean for class2
[c#3] (m3) !Latent variable mean for class3
[c#4] (m4) !Latent variable mean for class4
[c#5] (m5) !Latent variable mean for class5
[c#6] (m6) !Latent variable mean for class6
[c#7] (m7) !Latent variable mean for class7
[c#8] (m8) !Latent variable mean for class8
[c#9] (m9) !Latent variable mean for class9
[c#10] (m10) !Latent variable mean for class10
[c#11] (m11) !Latent variable mean for class11
[c#12] (m12) !Latent variable mean for class12
[c#13] (m13) !Latent variable mean for class13
[c#14] (m14) !Latent variable mean for class14
[c#15] (m15) !Latent variable mean for class15
[c#16] (m16) !Latent variable mean for class16
```

```
[c#17] (m17) !Latent variable mean for class17
[c#18] (m18) !Latent variable mean for class18
[c#19] (m19) !Latent variable mean for class19
[c#20] (m20) !Latent variable mean for class20
[c#21] (m21) !Latent variable mean for class21
[c#22] (m22) !Latent variable mean for class22
[c#23] (m23) !Latent variable mean for class23
[c#24] (m24) !Latent variable mean for class24
[c#25] (m25) !Latent variable mean for class25
[c#26] (m26) !Latent variable mean for class26
!=============================================================
=======

%c#1%    ![000]
[I1$1]    (T1_1);
[I2$1]    (T2_1);
[I3$1]    (T3_1);
[I4$1]    (T4_1);
[I5$1]    (T5_1);
[I6$1]    (T6_1);
[I7$1]    (T7_1);
[I8$1]    (T8_1);
[I9$1]    (T9_1);
[I10$1]    (T10_1);
[I11$1]    (T11_1);
[I12$1]    (T12_1);
[I13$1]    (T13_1);
[I14$1]    (T14_1);
[I15$1]    (T15_1);
[I16$1]    (T16_1);
[I17$1]    (T17_1);
[I18$1]    (T18_1);
[I19$1]    (T19_1);
[I20$1]    (T20_1);
[I21$1]    (T21_1);
[I22$1]    (T22_1);
[I23$1]    (T23_1);
[I24$1]    (T24_1);
!=============================================================
=======
%c#2%    ![001]
[I1$1]    (T1_1);
[I2$1]    (T2_1);
[I3$1]    (T3_1);
[I4$1]    (T4_1);
[I5$1]    (T5_1);
```

```
[I6$1]    (T6_1);
[I7$1]    (T7_1);
[I8$1]    (T8_1);
[I9$1]    (T9_1);
[I10$1]    (T10_1);
[I11$1]    (T11_1);
[I12$1]    (T12_1);
[I13$1]    (T13_1);
[I14$1]    (T14_1);
[I15$1]    (T15_1);
[I16$1]    (T16_1);
[I17$1]    (T17_2);
[I18$1]    (T18_2);
[I19$1]    (T19_2);
[I20$1]    (T20_2);
[I21$1]    (T21_2);
[I22$1]    (T22_2);
[I23$1]    (T23_2);
[I24$1]    (T24_2);
!=====================================================================
=======
%c#3%   ![002]
[I1$1]    (T1_1);
[I2$1]    (T2_1);
[I3$1]    (T3_1);
[I4$1]    (T4_1);
[I5$1]    (T5_1);
[I6$1]    (T6_1);
[I7$1]    (T7_1);
[I8$1]    (T8_1);
[I9$1]    (T9_1);
[I10$1]    (T10_1);
[I11$1]    (T11_1);
[I12$1]    (T12_1);
[I13$1]    (T13_1);
[I14$1]    (T14_1);
[I15$1]    (T15_1);
[I16$1]    (T16_1);
[I17$1]    (T17_3);
[I18$1]    (T18_3);
[I19$1]    (T19_3);
[I20$1]    (T20_3);
[I21$1]    (T21_3);
[I22$1]    (T22_3);
[I23$1]    (T23_3);
[I24$1]    (T24_3);
```

```
!=====================================================================
=======
%c#4%   ![010]
[I1$1]    (T1_1);
[I2$1]    (T2_1);
[I3$1]    (T3_1);
[I4$1]    (T4_1);
[I5$1]    (T5_1);
[I6$1]    (T6_1);
[I7$1]    (T7_1);
[I8$1]    (T8_1);
[I9$1]    (T9_2);
[I10$1]    (T10_2);
[I11$1]    (T11_2);
[I12$1]    (T12_2);
[I13$1]    (T13_2);
[I14$1]    (T14_2);
[I15$1]    (T15_2);
[I16$1]    (T16_2);
[I17$1]    (T17_1);
[I18$1]    (T18_1);
[I19$1]    (T19_1);
[I20$1]    (T20_1);
[I21$1]    (T21_1);
[I22$1]    (T22_1);
[I23$1]    (T23_1);
[I24$1]    (T24_1);
!=====================================================================
=======
%c#5%   ![011]
[I1$1]    (T1_1);
[I2$1]    (T2_1);
[I3$1]    (T3_1);
[I4$1]    (T4_1);
[I5$1]    (T5_1);
[I6$1]    (T6_1);
[I7$1]    (T7_1);
[I8$1]    (T8_1);
[I9$1]    (T9_2);
[I10$1]    (T10_2);
[I11$1]    (T11_2);
[I12$1]    (T12_2);
[I13$1]    (T13_2);
[I14$1]    (T14_2);
[I15$1]    (T15_2);
[I16$1]    (T16_2);
```

```
[I17$1]    (T17_2);
[I18$1]    (T18_2);
[I19$1]    (T19_2);
[I20$1]    (T20_2);
[I21$1]    (T21_2);
[I22$1]    (T22_2);
[I23$1]    (T23_2);
[I24$1]    (T24_2);
!=====================================================================
=======
%c#6%    ![012]
[I1$1]    (T1_1);
[I2$1]    (T2_1);
[I3$1]    (T3_1);
[I4$1]    (T4_1);
[I5$1]    (T5_1);
[I6$1]    (T6_1);
[I7$1]    (T7_1);
[I8$1]    (T8_1);
[I9$1]    (T9_2);
[I10$1]    (T10_2);
[I11$1]    (T11_2);
[I12$1]    (T12_2);
[I13$1]    (T13_2);
[I14$1]    (T14_2);
[I15$1]    (T15_2);
[I16$1]    (T16_2);
[I17$1]    (T17_3);
[I18$1]    (T18_3);
[I19$1]    (T19_3);
[I20$1]    (T20_3);
[I21$1]    (T21_3);
[I22$1]    (T22_3);
[I23$1]    (T23_3);
[I24$1]    (T24_3);
!=====================================================================
=======
%c#7%    ![020]
[I1$1]    (T1_1);
[I2$1]    (T2_1);
[I3$1]    (T3_1);
[I4$1]    (T4_1);
[I5$1]    (T5_1);
[I6$1]    (T6_1);
[I7$1]    (T7_1);
[I8$1]    (T8_1);
```

```
[I9$1]    (T9_3);
[I10$1]    (T10_3);
[I11$1]    (T11_3);
[I12$1]    (T12_3);
[I13$1]    (T13_3);
[I14$1]    (T14_3);
[I15$1]    (T15_3);
[I16$1]    (T16_3);
[I17$1]    (T17_1);
[I18$1]    (T18_1);
[I19$1]    (T19_1);
[I20$1]    (T20_1);
[I21$1]    (T21_1);
[I22$1]    (T22_1);
[I23$1]    (T23_1);
[I24$1]    (T24_1);
!=====================================================================
=======
%c#8%   ![021]
[I1$1]    (T1_1);
[I2$1]    (T2_1);
[I3$1]    (T3_1);
[I4$1]    (T4_1);
[I5$1]    (T5_1);
[I6$1]    (T6_1);
[I7$1]    (T7_1);
[I8$1]    (T8_1);
[I9$1]    (T9_3);
[I10$1]    (T10_3);
[I11$1]    (T11_3);
[I12$1]    (T12_3);
[I13$1]    (T13_3);
[I14$1]    (T14_3);
[I15$1]    (T15_3);
[I16$1]    (T16_3);
[I17$1]    (T17_2);
[I18$1]    (T18_2);
[I19$1]    (T19_2);
[I20$1]    (T20_2);
[I21$1]    (T21_2);
[I22$1]    (T22_2);
[I23$1]    (T23_2);
[I24$1]    (T24_2);
!=====================================================================
=======
%c#9%   ![022]
```

```
[I1$1]    (T1_1);
[I2$1]    (T2_1);
[I3$1]    (T3_1);
[I4$1]    (T4_1);
[I5$1]    (T5_1);
[I6$1]    (T6_1);
[I7$1]    (T7_1);
[I8$1]    (T8_1);
[I9$1]    (T9_3);
[I10$1]    (T10_3);
[I11$1]    (T11_3);
[I12$1]    (T12_3);
[I13$1]    (T13_3);
[I14$1]    (T14_3);
[I15$1]    (T15_3);
[I16$1]    (T16_3);
[I17$1]    (T17_3);
[I18$1]    (T18_3);
[I19$1]    (T19_3);
[I20$1]    (T20_3);
[I21$1]    (T21_3);
[I22$1]    (T22_3);
[I23$1]    (T23_3);
[I24$1]    (T24_3);
!=====================================================================
=======
%c#10%    ![100]
[I1$1]    (T1_2);
[I2$1]    (T2_2);
[I3$1]    (T3_2);
[I4$1]    (T4_2);
[I5$1]    (T5_2);
[I6$1]    (T6_2);
[I7$1]    (T7_2);
[I8$1]    (T8_2);
[I9$1]    (T9_1);
[I10$1]    (T10_1);
[I11$1]    (T11_1);
[I12$1]    (T12_1);
[I13$1]    (T13_1);
[I14$1]    (T14_1);
[I15$1]    (T15_1);
[I16$1]    (T16_1);
[I17$1]    (T17_1);
[I18$1]    (T18_1);
[I19$1]    (T19_1);
```

```
[I20$1]    (T20_1);
[I21$1]    (T21_1);
[I22$1]    (T22_1);
[I23$1]    (T23_1);
[I24$1]    (T24_1);
!=====================================================================
=======
%c#11%    ![101]
[I1$1]    (T1_2);
[I2$1]    (T2_2);
[I3$1]    (T3_2);
[I4$1]    (T4_2);
[I5$1]    (T5_2);
[I6$1]    (T6_2);
[I7$1]    (T7_2);
[I8$1]    (T8_2);
[I9$1]    (T9_1);
[I10$1]    (T10_1);
[I11$1]    (T11_1);
[I12$1]    (T12_1);
[I13$1]    (T13_1);
[I14$1]    (T14_1);
[I15$1]    (T15_1);
[I16$1]    (T16_1);
[I17$1]    (T17_2);
[I18$1]    (T18_2);
[I19$1]    (T19_2);
[I20$1]    (T20_2);
[I21$1]    (T21_2);
[I22$1]    (T22_2);
[I23$1]    (T23_2);
[I24$1]    (T24_2);
!=====================================================================
=======
%c#12%    ![102]
[I1$1]    (T1_2);
[I2$1]    (T2_2);
[I3$1]    (T3_2);
[I4$1]    (T4_2);
[I5$1]    (T5_2);
[I6$1]    (T6_2);
[I7$1]    (T7_2);
[I8$1]    (T8_2);
[I9$1]    (T9_1);
[I10$1]    (T10_1);
[I11$1]    (T11_1);
```

```
[I12$1]    (T12_1);
[I13$1]    (T13_1);
[I14$1]    (T14_1);
[I15$1]    (T15_1);
[I16$1]    (T16_1);
[I17$1]    (T17_3);
[I18$1]    (T18_3);
[I19$1]    (T19_3);
[I20$1]    (T20_3);
[I21$1]    (T21_3);
[I22$1]    (T22_3);
[I23$1]    (T23_3);
[I24$1]    (T24_3);
!=====================================================================
=======
%c#13%    ![110]
[I1$1]    (T1_2);
[I2$1]    (T2_2);
[I3$1]    (T3_2);
[I4$1]    (T4_2);
[I5$1]    (T5_2);
[I6$1]    (T6_2);
[I7$1]    (T7_2);
[I8$1]    (T8_2);
[I9$1]    (T9_2);
[I10$1]    (T10_2);
[I11$1]    (T11_2);
[I12$1]    (T12_2);
[I13$1]    (T13_2);
[I14$1]    (T14_2);
[I15$1]    (T15_2);
[I16$1]    (T16_2);
[I17$1]    (T17_1);
[I18$1]    (T18_1);
[I19$1]    (T19_1);
[I20$1]    (T20_1);
[I21$1]    (T21_1);
[I22$1]    (T22_1);
[I23$1]    (T23_1);
[I24$1]    (T24_1);
!=====================================================================
=======
%c#14%    ![111]
[I1$1]    (T1_2);
[I2$1]    (T2_2);
[I3$1]    (T3_2);
```

```
[I4$1]    (T4_2);
[I5$1]    (T5_2);
[I6$1]    (T6_2);
[I7$1]    (T7_2);
[I8$1]    (T8_2);
[I9$1]    (T9_2);
[I10$1]    (T10_2);
[I11$1]    (T11_2);
[I12$1]    (T12_2);
[I13$1]    (T13_2);
[I14$1]    (T14_2);
[I15$1]    (T15_2);
[I16$1]    (T16_2);
[I17$1]    (T17_2);
[I18$1]    (T18_2);
[I19$1]    (T19_2);
[I20$1]    (T20_2);
[I21$1]    (T21_2);
[I22$1]    (T22_2);
[I23$1]    (T23_2);
[I24$1]    (T24_2);
!=====================================================================
=======
%c#15%    ![112]
[I1$1]    (T1_2);
[I2$1]    (T2_2);
[I3$1]    (T3_2);
[I4$1]    (T4_2);
[I5$1]    (T5_2);
[I6$1]    (T6_2);
[I7$1]    (T7_2);
[I8$1]    (T8_2);
[I9$1]    (T9_2);
[I10$1]    (T10_2);
[I11$1]    (T11_2);
[I12$1]    (T12_2);
[I13$1]    (T13_2);
[I14$1]    (T14_2);
[I15$1]    (T15_2);
[I16$1]    (T16_2);
[I17$1]    (T17_3);
[I18$1]    (T18_3);
[I19$1]    (T19_3);
[I20$1]    (T20_3);
[I21$1]    (T21_3);
[I22$1]    (T22_3);
```

```
[I23$1]    (T23_3);
[I24$1]    (T24_3);
!======================================================================
=======
%c#16%    ![120]
[I1$1]    (T1_2);
[I2$1]    (T2_2);
[I3$1]    (T3_2);
[I4$1]    (T4_2);
[I5$1]    (T5_2);
[I6$1]    (T6_2);
[I7$1]    (T7_2);
[I8$1]    (T8_2);
[I9$1]    (T9_3);
[I10$1]    (T10_3);
[I11$1]    (T11_3);
[I12$1]    (T12_3);
[I13$1]    (T13_3);
[I14$1]    (T14_3);
[I15$1]    (T15_3);
[I16$1]    (T16_3);
[I17$1]    (T17_1);
[I18$1]    (T18_1);
[I19$1]    (T19_1);
[I20$1]    (T20_1);
[I21$1]    (T21_1);
[I22$1]    (T22_1);
[I23$1]    (T23_1);
[I24$1]    (T24_1);
!======================================================================
=======
%c#17%    ![121]
[I1$1]    (T1_2);
[I2$1]    (T2_2);
[I3$1]    (T3_2);
[I4$1]    (T4_2);
[I5$1]    (T5_2);
[I6$1]    (T6_2);
[I7$1]    (T7_2);
[I8$1]    (T8_2);
[I9$1]    (T9_3);
[I10$1]    (T10_3);
[I11$1]    (T11_3);
[I12$1]    (T12_3);
[I13$1]    (T13_3);
[I14$1]    (T14_3);
```

```
[I15$1]    (T15_3);
[I16$1]    (T16_3);
[I17$1]    (T17_2);
[I18$1]    (T18_2);
[I19$1]    (T19_2);
[I20$1]    (T20_2);
[I21$1]    (T21_2);
[I22$1]    (T22_2);
[I23$1]    (T23_2);
[I24$1]    (T24_2);
!=================================================================
=======
%c#18%    ![122]
[I1$1]    (T1_2);
[I2$1]    (T2_2);
[I3$1]    (T3_2);
[I4$1]    (T4_2);
[I5$1]    (T5_2);
[I6$1]    (T6_2);
[I7$1]    (T7_2);
[I8$1]    (T8_2);
[I9$1]    (T9_3);
[I10$1]    (T10_3);
[I11$1]    (T11_3);
[I12$1]    (T12_3);
[I13$1]    (T13_3);
[I14$1]    (T14_3);
[I15$1]    (T15_3);
[I16$1]    (T16_3);
[I17$1]    (T17_3);
[I18$1]    (T18_3);
[I19$1]    (T19_3);
[I20$1]    (T20_3);
[I21$1]    (T21_3);
[I22$1]    (T22_3);
[I23$1]    (T23_3);
[I24$1]    (T24_3);
!=================================================================
=======
%c#19%    ![200]
[I1$1]    (T1_3);
[I2$1]    (T2_3);
[I3$1]    (T3_3);
[I4$1]    (T4_3);
[I5$1]    (T5_3);
[I6$1]    (T6_3);
```

```
[I7$1]    (T7_3);
[I8$1]    (T8_3);
[I9$1]    (T9_1);
[I10$1]    (T10_1);
[I11$1]    (T11_1);
[I12$1]    (T12_1);
[I13$1]    (T13_1);
[I14$1]    (T14_1);
[I15$1]    (T15_1);
[I16$1]    (T16_1);
[I17$1]    (T17_1);
[I18$1]    (T18_1);
[I19$1]    (T19_1);
[I20$1]    (T20_1);
[I21$1]    (T21_1);
[I22$1]    (T22_1);
[I23$1]    (T23_1);
[I24$1]    (T24_1);
!=====================================================================
=======
%c#20%    ![201]
[I1$1]    (T1_3);
[I2$1]    (T2_3);
[I3$1]    (T3_3);
[I4$1]    (T4_3);
[I5$1]    (T5_3);
[I6$1]    (T6_3);
[I7$1]    (T7_3);
[I8$1]    (T8_3);
[I9$1]    (T9_1);
[I10$1]    (T10_1);
[I11$1]    (T11_1);
[I12$1]    (T12_1);
[I13$1]    (T13_1);
[I14$1]    (T14_1);
[I15$1]    (T15_1);
[I16$1]    (T16_1);
[I17$1]    (T17_2);
[I18$1]    (T18_2);
[I19$1]    (T19_2);
[I20$1]    (T20_2);
[I21$1]    (T21_2);
[I22$1]    (T22_2);
[I23$1]    (T23_2);
[I24$1]    (T24_2);
```

```
!=====================================================================
=======
%c#21%   ![202]
[I1$1]    (T1_3);
[I2$1]    (T2_3);
[I3$1]    (T3_3);
[I4$1]    (T4_3);
[I5$1]    (T5_3);
[I6$1]    (T6_3);
[I7$1]    (T7_3);
[I8$1]    (T8_3);
[I9$1]    (T9_1);
[I10$1]    (T10_1);
[I11$1]    (T11_1);
[I12$1]    (T12_1);
[I13$1]    (T13_1);
[I14$1]    (T14_1);
[I15$1]    (T15_1);
[I16$1]    (T16_1);
[I17$1]    (T17_3);
[I18$1]    (T18_3);
[I19$1]    (T19_3);
[I20$1]    (T20_3);
[I21$1]    (T21_3);
[I22$1]    (T22_3);
[I23$1]    (T23_3);
[I24$1]    (T24_3);
!=====================================================================
=======
%c#22%   ![210]
[I1$1]    (T1_3);
[I2$1]    (T2_3);
[I3$1]    (T3_3);
[I4$1]    (T4_3);
[I5$1]    (T5_3);
[I6$1]    (T6_3);
[I7$1]    (T7_3);
[I8$1]    (T8_3);
[I9$1]    (T9_2);
[I10$1]    (T10_2);
[I11$1]    (T11_2);
[I12$1]    (T12_2);
[I13$1]    (T13_2);
[I14$1]    (T14_2);
[I15$1]    (T15_2);
[I16$1]    (T16_2);
```

```
[I17$1]    (T17_1);
[I18$1]    (T18_1);
[I19$1]    (T19_1);
[I20$1]    (T20_1);
[I21$1]    (T21_1);
[I22$1]    (T22_1);
[I23$1]    (T23_1);
[I24$1]    (T24_1);
!=====================================================================
=======
%c#23%   ![211]
[I1$1]    (T1_3);
[I2$1]    (T2_3);
[I3$1]    (T3_3);
[I4$1]    (T4_3);
[I5$1]    (T5_3);
[I6$1]    (T6_3);
[I7$1]    (T7_3);
[I8$1]    (T8_3);
[I9$1]    (T9_2);
[I10$1]    (T10_2);
[I11$1]    (T11_2);
[I12$1]    (T12_2);
[I13$1]    (T13_2);
[I14$1]    (T14_2);
[I15$1]    (T15_2);
[I16$1]    (T16_2);
[I17$1]    (T17_2);
[I18$1]    (T18_2);
[I19$1]    (T19_2);
[I20$1]    (T20_2);
[I21$1]    (T21_2);
[I22$1]    (T22_2);
[I23$1]    (T23_2);
[I24$1]    (T24_2);
!=====================================================================
=======
%c#24%   ![212]
[I1$1]    (T1_3);
[I2$1]    (T2_3);
[I3$1]    (T3_3);
[I4$1]    (T4_3);
[I5$1]    (T5_3);
[I6$1]    (T6_3);
[I7$1]    (T7_3);
[I8$1]    (T8_3);
```

```
[I9$1]    (T9_2);
[I10$1]    (T10_2);
[I11$1]    (T11_2);
[I12$1]    (T12_2);
[I13$1]    (T13_2);
[I14$1]    (T14_2);
[I15$1]    (T15_2);
[I16$1]    (T16_2);
[I17$1]    (T17_3);
[I18$1]    (T18_3);
[I19$1]    (T19_3);
[I20$1]    (T20_3);
[I21$1]    (T21_3);
[I22$1]    (T22_3);
[I23$1]    (T23_3);
[I24$1]    (T24_3);
!==================================================================
=======
%c#25%   ![220]
[I1$1]    (T1_3);
[I2$1]    (T2_3);
[I3$1]    (T3_3);
[I4$1]    (T4_3);
[I5$1]    (T5_3);
[I6$1]    (T6_3);
[I7$1]    (T7_3);
[I8$1]    (T8_3);
[I9$1]    (T9_3);
[I10$1]    (T10_3);
[I11$1]    (T11_3);
[I12$1]    (T12_3);
[I13$1]    (T13_3);
[I14$1]    (T14_3);
[I15$1]    (T15_3);
[I16$1]    (T16_3);
[I17$1]    (T17_1);
[I18$1]    (T18_1);
[I19$1]    (T19_1);
[I20$1]    (T20_1);
[I21$1]    (T21_1);
[I22$1]    (T22_1);
[I23$1]    (T23_1);
[I24$1]    (T24_1);
!==================================================================
=======
%c#26%   ![221]
```

```
[I1$1]    (T1_3);
[I2$1]    (T2_3);
[I3$1]    (T3_3);
[I4$1]    (T4_3);
[I5$1]    (T5_3);
[I6$1]    (T6_3);
[I7$1]    (T7_3);
[I8$1]    (T8_3);
[I9$1]    (T9_3);
[I10$1]    (T10_3);
[I11$1]    (T11_3);
[I12$1]    (T12_3);
[I13$1]    (T13_3);
[I14$1]    (T14_3);
[I15$1]    (T15_3);
[I16$1]    (T16_3);
[I17$1]    (T17_2);
[I18$1]    (T18_2);
[I19$1]    (T19_2);
[I20$1]    (T20_2);
[I21$1]    (T21_2);
[I22$1]    (T22_2);
[I23$1]    (T23_2);
[I24$1]    (T24_2);
!=====================================================================
=======
%c#27%    ![222]
[I1$1]    (T1_3);
[I2$1]    (T2_3);
[I3$1]    (T3_3);
[I4$1]    (T4_3);
[I5$1]    (T5_3);
[I6$1]    (T6_3);
[I7$1]    (T7_3);
[I8$1]    (T8_3);
[I9$1]    (T9_3);
[I10$1]    (T10_3);
[I11$1]    (T11_3);
[I12$1]    (T12_3);
[I13$1]    (T13_3);
[I14$1]    (T14_3);
[I15$1]    (T15_3);
[I16$1]    (T16_3);
[I17$1]    (T17_3);
[I18$1]    (T18_3);
[I19$1]    (T19_3);
```

```
[I20$1]     (T20_3);
[I21$1]     (T21_3);
[I22$1]     (T22_3);
[I23$1]     (T23_3);
[I24$1]     (T24_3);

!================================================================
=======

MODEL CONSTRAINT:  ! Used to define LCDM parameters
  ! Mplus uses P(X=0) rather than P(X=1) so multiply by -1
NEW(G_0 G_11_1 G_11_2 G_12_1 G_12_2 G_13_1 G_13_2
G_212_11 G_212_12 G_212_21 G_212_22 G_213_11 G_213_12
G_213_21 G_213_22 G_223_11 G_223_12 G_223_21 G_223_22);
G_0 = -(G_11_1+G_11_2+G_12_1+G_12_2+G_13_1+G_13_2+G_212_11+
      G_212_12+G_212_21+G_212_22+G_213_11+G_213_12+G_213_21+
      G_213_22+G_223_11+G_223_12+G_223_21+G_223_22);
m1 = G_0;
m2 = G_0+G_13_1;
m3 = G_0+G_13_1+G_13_2;
m4 = G_0+G_12_1;
m5 = G_0+G_12_1+G_13_1+G_223_11;
m6 = G_0+G_12_1+G_13_1+G_13_2+G_223_11+G_223_12;
m7 = G_0+G_12_1+G_12_2;
m8 = G_0+G_12_1+G_12_2+G_13_1+G_223_11+G_223_21;
m9 = G_0+G_12_1+G_12_2+G_13_1+G_13_2+G_223_11+G_223_12+
     G_223_21+G_223_22;
m10 = G_0+G_11_1;
m11 = G_0+G_11_1+G_13_1+G_213_11;
m12 = G_0+G_11_1+G_13_1+G_13_2+G_213_11+G_213_12;
m13 = G_0+G_11_1+G_12_1+G_212_11;
m14 = G_0+G_11_1+G_12_1+G_13_1+G_212_11+G_213_11+G_223_11;
m15 = G_0+G_11_1+G_12_1+G_13_1+G_212_11+G_213_11+G_223_11+
      G_223_12+G_213_12;
m16 = G_0+G_11_1+G_12_1+G_12_2+G_212_11+G_212_12;
m17 = G_0+G_11_1+G_12_1+G_12_2+G_13_1+G_212_11+G_212_12+
      G_213_11+G_223_11+G_223_21;
m18 = G_0+G_11_1+G_12_1+G_12_2+G_13_1+G_13_2+G_212_11+G_212_12+
      G_213_11+G_213_12+G_223_11+G_223_21+G_223_12+G_223_22;
m19 = G_0+G_11_1+G_11_2;
m20 = G_0+G_11_1+G_11_2+G_13_1+G_213_11+G_213_21;
m21 = G_0+G_11_1+G_11_2+G_13_1+G_13_2+G_213_11+G_213_21+
      G_213_12+G_213_22;
m22 = G_0+G_11_1+G_11_2+G_12_1+G_212_11+G_212_21;
m23 = G_0+G_11_1+G_11_2+G_12_1+G_13_1+G_212_11+G_212_21+
      G_213_11+G_213_21+G_223_11;
```

```
m24 = G_0+G_11_1+G_11_2+G_12_1+G_13_1+G_13_2+G_212_11+G_212_21+
    G_213_11+G_213_21+G_223_11+G_223_12;
m25 = G_0+G_11_1+G_11_2+G_12_1+G_12_2+G_212_11+G_212_12+G_212_21+
    G_212_22;
m26 = G_0+G_11_1+G_11_2+G_12_1+G_12_2+G_13_1+G_212_11+G_212_12+
    G_212_21+G_212_22+G_213_11+G_213_21+G_223_11+G_223_21;

! Item 1: Define LCDM parameters present for item 1
NEW(L1_0 L1_11_1 L1_11_2);
T1_1=-(L1_0);
T1_2=-(L1_0+L1_11_1);
T1_3=-(L1_0+L1_11_1+L1_11_2);
L1_0>-10; L1_0<10;
! Main effect order constraints
L1_11_1>0; L1_11_2>0;
L1_11_1<10; L1_11_2<10;

! Item 2: Define LCDM parameters present for item 2
NEW(L2_0 L2_11_1 L2_11_2);
T2_1=-(L2_0);
T2_2=-(L2_0+L2_11_1);
T2_3=-(L2_0+L2_11_1+L2_11_2);
L2_0>-10; L2_0<10;
! Main effect order constraints
L2_11_1>0; L2_11_2>0;
L2_11_1<10; L2_11_2<10;


! Item 3: Define LCDM parameters present for item 3
NEW(L3_0 L3_11_1 L3_11_2);
T3_1=-(L3_0);
T3_2=-(L3_0+L3_11_1);
T3_3=-(L3_0+L3_11_1+L3_11_2);
L3_0>-10; L3_0<10;
! Main effect order constraints
L3_11_1>0; L3_11_2>0;
L3_11_1<10; L3_11_2<10;

! Item 4: Define LCDM parameters present for item 4
NEW(L4_0 L4_11_1 L4_11_2);
T4_1=-(L4_0);
T4_2=-(L4_0+L4_11_1);
T4_3=-(L4_0+L4_11_1+L4_11_2);
L4_0>-10; L4_0<10;
! Main effect order constraints
L4_11_1>0; L4_11_2>0;
```

```
L4_11_1<10; L4_11_2<10;

! Item 5: Define LCDM parameters present for item 5
NEW(L5_0 L5_11_1 L5_11_2);
T5_1=-(L5_0);
T5_2=-(L5_0+L5_11_1);
T5_3=-(L5_0+L5_11_1+L5_11_2);
L5_0>-10; L5_0<10;
! Main effect order constraints
L5_11_1>0; L5_11_2>0;
L5_11_1<10; L5_11_2<10;

! Item 6: Define LCDM parameters present for item 6
NEW(L6_0 L6_11_1 L6_11_2);
T6_1=-(L6_0);
T6_2=-(L6_0+L6_11_1);
T6_3=-(L6_0+L6_11_1+L6_11_2);
L6_0>-10; L6_0<10;
! Main effect order constraints
L6_11_1>0; L6_11_2>0;
L6_11_1<10; L6_11_2<10;

! Item 7: Define LCDM parameters present for item 7
NEW(L7_0 L7_11_1 L7_11_2);
T7_1=-(L7_0);
T7_2=-(L7_0+L7_11_1);
T7_3=-(L7_0+L7_11_1+L7_11_2);
L7_0>-10; L7_0<10;
! Main effect order constraints
L7_11_1>0; L7_11_2>0;
L7_11_1<10; L7_11_2<10;

! Item 8: Define LCDM parameters present for item 8
NEW(L8_0 L8_11_1 L8_11_2);
T8_1=-(L8_0);
T8_2=-(L8_0+L8_11_1);
T8_3=-(L8_0+L8_11_1+L8_11_2);
L8_0>-10; L8_0<10;
! Main effect order constraints
L8_11_1>0; L8_11_2>0;
L8_11_1<10; L8_11_2<10;

! Item 9: Define LCDM parameters present for item 9
NEW(L9_0 L9_12_1 L9_12_2);
T9_1=-(L9_0);
T9_2=-(L9_0+L9_12_1);
```

```
T9_3=-(L9_0+L9_12_1+L9_12_2);
L9_0>-10; L9_0<10;
! Main effect order constraints
L9_12_1>0; L9_12_2>0;
L9_12_1<10; L9_12_2<10;

! Item 10: Define LCDM parameters present for item 10
NEW(L10_0 L10_12_1 L10_12_2);
T10_1=-(L10_0);
T10_2=-(L10_0+L10_12_1);
T10_3=-(L10_0+L10_12_1+L10_12_2);
L10_0>-10; L10_0<10;
! Main effect order constraints
L10_12_1>0; L10_12_2>0;
L10_12_1<10; L10_12_2<10;

! Item 11: Define LCDM parameters present for item 11
NEW(L11_0 L11_12_1 L11_12_2);
T11_1=-(L11_0);
T11_2=-(L11_0+L11_12_1);
T11_3=-(L11_0+L11_12_1+L11_12_2);
L11_0>-10; L11_0<10;
! Main effect order constraints
L11_12_1>0; L11_12_2>0;
L11_12_1<10; L11_12_2<10;

! Item 12: Define LCDM parameters present for item 12
NEW(L12_0 L12_12_1 L12_12_2);
T12_1=-(L12_0);
T12_2=-(L12_0+L12_12_1);
T12_3=-(L12_0+L12_12_1+L12_12_2);
L12_0>-10; L12_0<10;
! Main effect order constraints
L12_12_1>0; L12_12_2>0;
L12_12_1<10; L12_12_2<10;

! Item 13: Define LCDM parameters present for item 13
NEW(L13_0 L13_12_1 L13_12_2);
T13_1=-(L13_0);
T13_2=-(L13_0+L13_12_1);
T13_3=-(L13_0+L13_12_1+L13_12_2);
L13_0>-10; L13_0<10;
! Main effect order constraints
L13_12_1>0; L13_12_2>0;
L13_12_1<10; L13_12_2<10;
```

```
! Item 14: Define LCDM parameters present for item 14
NEW(L14_0 L14_12_1 L14_12_2);
T14_1=-(L14_0);
T14_2=-(L14_0+L14_12_1);
T14_3=-(L14_0+L14_12_1+L14_12_2);
L14_0>-10; L14_0<10;
! Main effect order constraints
L14_12_1>0; L14_12_2>0;
L14_12_1<10; L14_12_2<10;

! Item 15: Define LCDM parameters present for item 15
NEW(L15_0 L15_12_1 L15_12_2);
T15_1=-(L15_0);
T15_2=-(L15_0+L15_12_1);
T15_3=-(L15_0+L15_12_1+L15_12_2);
L15_0>-10; L15_0<10;
! Main effect order constraints
L15_12_1>0; L15_12_2>0;
L15_12_1<10; L15_12_2<10;

! Item 16: Define LCDM parameters present for item 16
NEW(L16_0 L16_12_1 L16_12_2);
T16_1=-(L16_0);
T16_2=-(L16_0+L16_12_1);
T16_3=-(L16_0+L16_12_1+L16_12_2);
L16_0>-10; L16_0<10;
! Main effect order constraints
L16_12_1>0; L16_12_2>0;
L16_12_1<10; L16_12_2<10;

! Item 17: Define LCDM parameters present for item 17
NEW(L17_0 L17_13_1 L17_13_2);
T17_1=-(L17_0);
T17_2=-(L17_0+L17_13_1);
T17_3=-(L17_0+L17_13_1+L17_13_2);
L17_0>-10; L17_0<10;
! Main effect order constraints
L17_13_1>0; L17_13_2>0;
L17_13_1<10; L17_13_2<10;

! Item 18: Define LCDM parameters present for item 18
NEW(L18_0 L18_13_1 L18_13_2);
T18_1=-(L18_0);
T18_2=-(L18_0+L18_13_1);
T18_3=-(L18_0+L18_13_1+L18_13_2);
L18_0>-10; L18_0<10;
```

```
! Main effect order constraints
L18_13_1>0; L18_13_2>0;
L18_13_1<10; L18_13_2<10;

! Item 19: Define LCDM parameters present for item 19
NEW(L19_0 L19_13_1 L19_13_2);
T19_1=-(L19_0);
T19_2=-(L19_0+L19_13_1);
T19_3=-(L19_0+L19_13_1+L19_13_2);
L19_0>-10; L19_0<10;
! Main effect order constraints
L19_13_1>0; L19_13_2>0;
L19_13_1<10; L19_13_2<10;

! Item 20: Define LCDM parameters present for item 20
NEW(L20_0 L20_13_1 L20_13_2);
T20_1=-(L20_0);
T20_2=-(L20_0+L20_13_1);
T20_3=-(L20_0+L20_13_1+L20_13_2);
L20_0>-10; L20_0<10;
! Main effect order constraints
L20_13_1>0; L20_13_2>0;
L20_13_1<10; L20_13_2<10;

! Item 21: Define LCDM parameters present for item 21
NEW(L21_0 L21_13_1 L21_13_2);
T21_1=-(L21_0);
T21_2=-(L21_0+L21_13_1);
T21_3=-(L21_0+L21_13_1+L21_13_2);
L21_0>-10; L21_0<10;
! Main effect order constraints
L21_13_1>0; L21_13_2>0;
L21_13_1<10; L21_13_2<10;

! Item 22: Define LCDM parameters present for item 22
NEW(L22_0 L22_13_1 L22_13_2);
T22_1=-(L22_0);
T22_2=-(L22_0+L22_13_1);
T22_3=-(L22_0+L22_13_1+L22_13_2);
L22_0>-10; L22_0<10;
! Main effect order constraints
L22_13_1>0; L22_13_2>0;
L22_13_1<10; L22_13_2<10;

! Item 23: Define LCDM parameters present for item 23
NEW(L23_0 L23_13_1 L23_13_2);
```

```
T23_1=-(L23_0);
T23_2=-(L23_0+L23_13_1);
T23_3=-(L23_0+L23_13_1+L23_13_2);
L23_0>-10; L23_0<10;
! Main effect order constraints
L23_13_1>0; L23_13_2>0;
L23_13_1<10; L23_13_2<10;

! Item 24: Define LCDM parameters present for item 24
NEW(L24_0 L24_13_1 L24_13_2);
T24_1=-(L24_0);
T24_2=-(L24_0+L24_13_1);
T24_3=-(L24_0+L24_13_1+L24_13_2);
L24_0>-10; L24_0<10;
! Main effect order constraints
L24_13_1>0; L24_13_2>0;
L24_13_1<10; L24_13_2<10;

OUTPUT:
   TECH10;  ! Request additional model fit statistics

SAVEDATA: ! Format, name of posterior probabilities of class membership file
   FORMAT = F10.5;
   FILE = respondents1.dat;
   SAVE = CPROBABILITIES;
```