Evaluation of Attribute Structure for a Dynamic Assessment Using the Loglinear Cognitive Diagnosis Model and Bayesian Networks

By

Feng Chen
Ph. D, The University of Kansas, 2018
M.S., The University of Kansas, 2011
B.A., Jilin Huaqiao University of Foreign Languages, 2009

Submitted to the graduate degree program in Department of Educational Psychology and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

_____

Chair: Jonathan Templin

_____

David Hansen

_____

Vicki Peyton

_____

Bruce Frey

_____

Lesa Hoffman

Date Defended: 28 September 2018

The dissertation committee for Feng Chen certifies that this is the approved version of the following dissertation:

Evaluation of Attribute Structure for a Dynamic Assessment Using the Loglinear Cognitive Diagnosis Model and Bayesian Networks

_____

Chair: Jonathan Templin

Date Approved: 28 September 2018

Abstract

In the field of assessment, the construction of a test is critical in matters of pinpointing the use and purpose of the test, the models to be used to generate results, and the inferences that can be made from the test results. Although an attribute map is not necessary to construct a good assessment, a series of well-delineated sets of attributes and a set of well-developed items written to the attributes are essential. As the practitioners of multidimensional test and diagnostic classification models (DCMs) grow, it is important to examine the underlying structural models of attributes within a test. The dissertation seeks to examine the possible structural models of the attributes, using both real data, the Diagnosing Teachers' Multiplicative Reasoning assessment (Bradshaw et al., 2014) and simulated data in the framework of Loglinear Cognitive Diagnosis Model (LCDM) and Bayesian Networks. Additionally, this research explores the methodology for possible attribute structures that maximizes the impact of the map structure to the implementation and development of a diagnostic assessment. Results from the analysis indicate that the selection of attribute structure can have some implications for attribute parameter estimates and student mastery classifications. The findings also show that sample size and test length have more impact on item level parameter estimates. In addition, the results demonstrated that LCDM integrated with Bayesian Networks is a feasible methodology to detect attribute hierarchy, and thus is a practical choice for multidimensional test scoring.

Acknowledgements

First and foremost, I would like to thank the Lord for carrying me through the journey of doctoral research. I am a recipient of so much generosity and kindness from the people around me, and I am truly grateful. I thank my chair, Jonathan Templin, for his mentorship, guidance, support and encouragement through the process. Jonathan, thank you, I would never have finished my dissertation without your guidance and support. Thank you for opening the door of diagnostic models and leading me into the diagnostic assessment field. I also greatly appreciate the time you have spent mentoring me, meeting with me in-person, through the phone, and answering my numerous questions via emails. I would also like to thank the rest of my committee: Bruce Frey, Vicki Peyton, Lesa Hoffman and David Hansen, for your support during the course of my dissertation writing and defending.

I also must thank my UMKC-IHD family and coworkers, specifically my superior George Gotto, and my coworker Megan Snyder. George, thank you for motivating me to finish my degree, and opening the door to so many opportunities for me to grow. You always have so much confidence and faith in me, even when I am questioning myself. I am so grateful that you are making every way to support my growth and career development. Many thanks to you, George. Megan, thank you for spending time on proofreading my dissertation. I know it was not easy for you during that couple of weeks, especially when Will is not sleeping at night. I really appreciate that you offered to help and tried to fit into my schedule. I cannot express how much appreciation I have toward your kindness and support.

Last but not least, I would like to thank my wonderful husband Michael. Without you I would not have stuck through. Thank you for always standing by my side supporting and listening whenever I am frustrated with my dissertation progress and programming issues. This is

as much your accomplishment as it is mine. Thank you for being in my life. I love you! I also must thank my parents, Sailiang Chen and Xiuhua Wang, who are caring and encouraging throughout the years of my doctoral study. Mom and Dad, thank you for praying for me constantly and having unquestioning belief in me. You are the true role model to me. To my in-laws, John and Holly Chiang, thank you for welcoming me into your family, and taking care of our dog Zoey when Michael and I need a get-away. To my dog Zoey, who can't read this but expressed her encouragement by giving me warm snuggles. Finally, I would like to thank my Lawrence friends: Li Chen-Bourke, Xuan Gu, and Yuping Hsu, who are always there for me when I need to fall back.

Table of Contents

Abstract iii
Acknowledgements iv
CHAPTER ONE 1
Introduction 1
CHAPTER TWO 7
Literature Review 7
Diagnostic Classification Models (DCMs) 7
Elements of Diagnostic Classification Models 8
Loglinear Cognitive Diagnosis Model (LCDM) 9
Comparison and Contrast of DCMs and IRT 11
Application of LCDM in Bayesian Networks 12
CHAPTER THREE 16
Methodology 16
Data Analysis 24
Evaluation criteria 25
Simulation Results 27
The Measures of Convergence for the MCMC Algorithm in the Bayesian Network 27
Deviance 30
Item Level Results 32
Results for Item Intercept Parameters 32
Results for Item Main Effect Parameters 33
Results for Item Interaction Parameters 34
Attribute Level Results 37
Results for Attribute Intercept Parameter 37
Results for Attribute Main Effect Parameter 41
Results for Attribute Interaction Parameter 42
Student Mastery Classification Accuracy 44
CHAPTER FIVE 48
Real Data Analysis Results 48
Model Fit and Model Convergence 49
Results for Attribute Parameters 66
Results for Mastery Classification 70

CHAPTER SIX 73
Discussion and Conclusion 73
Impact of Attribute Structure 73
Methodological Contributions 76
Conclusion 78
References 79
Appendix A: JAGS Code for Data Estimation 84

CHAPTER ONE

Introduction

In the field of assessment, the construction of a test is critical in matters of pinpointing the use and purpose of the test, the models to be used to generate results, and the inferences that can be made from the test results. Empirically, there have been numerous research studies on building and evaluating a test with the anticipation of using unidimensional models, such as item response theory (IRT) models. However, only a handful of studies focus on the topic of building a multi-dimensional test, due to reasons such as model-data fit, budget, the expertise of item development, and so forth. Therefore, findings based on real data with simulation study would be informative to further research and practice on multidimensional tests. With the lack of research in view, the purpose of this study is to explore existing research on multidimensional tests based on diagnostic classification models (DCMs) to a structural modeling of attributes within the test.

In spite of the controversies around multidimensional tests, the challenges in test development and model data retrofit, Bradshaw, Izsak, Templin, and Jacobson (2014) successfully built a multidimensional test within the diagnostic classification model framework. The test examines middle grades teachers' understanding of multiplication and division fraction arithmetic using four attributes: *Referent Units (RU), Partitioning and Iterating (PI), Appropriateness (APP),* and *Multiplicative Comparison (MC).* In this case, the term "attribute" denotes a binary latent trait that is not able to be measured directly and has to be investigated through the expressive features from the latent trait (e.g., knowledge, skill, ability, and etc.).

In the study by Bradshaw, et al. (2014), the attributes refer to teachers' knowledge of fractional reasoning. As mentioned above, four attributes are involved: *RU, PI, APP,* and *MC. RU* refers to *Referent Units,* which has to do with identifying fractional numbers. The second

attribute, *Partitioning and Iterating*, combines partitioning a whole number into equal pieces and iterating the meaning of it. The third attribute, *Appropriateness*, refers to identifying the proper operations needed to solve the problem. The last attribute, *Multiplicative Comparison*, requires making comparisons between or even among two or more values using the knowledge of multiplication (Bradshaw, Izsak, Templin, & Jacobson, 2014).

The findings pointed out that a multidimensional test is possible in design and implementation, and that DCMs as the tools are feasible to enable diagnostically reliable interpretations of teachers' abilities with fractional arithmetic (Bradshaw, Izsak, Templin, & Jacobson, 2014). The test developed by Bradshaw, et al. (2014) was built to measure teachers' knowledge of fractional reasoning, and the vertical bars in Figure 1.1 represents the proportion of teachers that have mastered each of the four attributes. From the graph, attribute *APP* was mastered by over 60% of the teachers, followed by *MC* and *PI*, and *RU*. However, attribute *RU* turns out to be the least mastered attribute by teachers nationwide, with a percentage of roughly 30%. Table 1.1 presents the proportion of teachers who are classified with each attribute pattern of mastery, according to the four attributes *RU, PI, APP,* and *MC*. As indicated in the table, a proportion of 25% teachers have a mastery of all four attributes, followed by 22% of teachers' have mastered none of the four attributes, and 15% have a mastery of three attributes – *APP*, *PI*, and *MC*, but not *RU*. Moreover, the proportion of teachers' mastery of an attribute pattern with *RU* been mastered is very low. For instance, just 1% teachers have a mastery of an attribute pattern with only *MC* and *RU* been mastered, and roughly 2% teachers have a mastery profile of an attribute pattern with *PI, MC,* and *RU* been mastered. With that being said, in an attribute pattern where *APP, PI,* and *MC* were all mastered (15% of teachers), the proportion of teachers' mastery of all four attributes (including *RU* been mastered) increased from 15% to 25%. The

question is why is there such a big leap in the proportion of teachers' mastery profile of attribute patterns when *APP*, *PI*, and *MC* were all mastered.

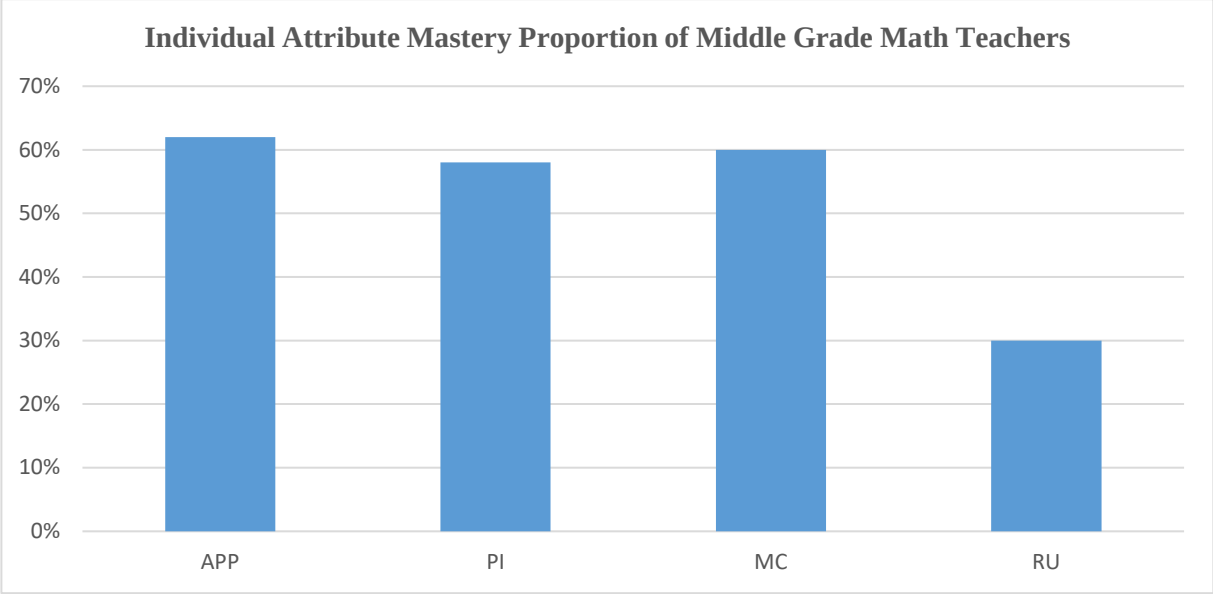

Figure 1.1. Individual Attribute Mastery Proportion. Referenced from "Diagnosing Teachers' Understandings of Rational Numbers: Building a Multidimensional Test Within the Diagnostic Classification Framework," by L. Bradshaw, A. Izsak, J. Templin and E. Jacobson, 2014, *Educational measurement: Issues and practice*, *33*(1), p. 12. Copyright 2013 by the National Council on Measurement in Education.

Table 1.1

Attribute Pattern Mastery Proportion

APP	PI	MC	RU	Proportion
0	0	0	0	22%
0	0	1	0	8%
1	0	0	0	6%
1	0	1	0	9%
0	1	0	0	5%
0	1	1	0	3%
1	1	0	0	7%
1	1	1	0	15%
0	0	0	1	1%
0	0	1	1	1%
1	0	0	1	1%
1	0	1	1	3%

0	1	0	1	1%
0	1	1	1	2%
1	1	0	1	3%
1	1	1	1	25%

Although the mastery pattern and proportion of mastery profile of the four attributes has been examined, the hierarchies among the attributes were not explored in Bradshaw, et al.'s study. In other words, teachers who have a better conception of the first attribute (*APP*) tend to also master the second (*PI*), third (*MC*), and even the fourth attribute (*RU*; see Table 1.1). When considering the hierarchies among the four attributes, a possible attribute map can be structured. A map is essentially a diagram or network of certain points, attributes in this case, and consists of attributes as dots and structural paths as connections. The attribute structure could look like this:

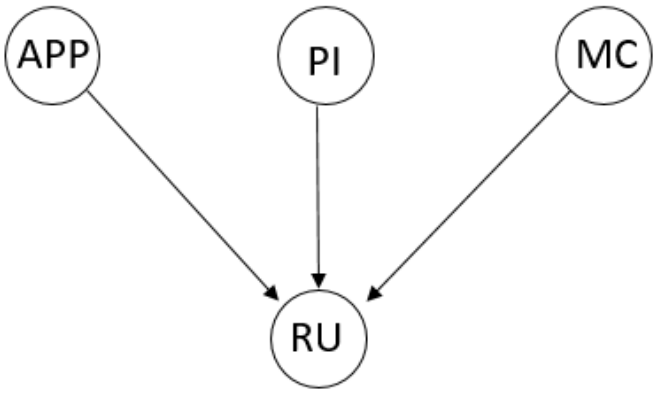

Figure 1.2. An Example of the attribute map based on the four attributes.

An example of the attribute map, inspired by Bradshaw, et al.'s data, is presented in Figure 1.2. Assuming three attributes, *APP, PI, MC,* are the foundational knowledge/skills, they progress to the target skill *RU*. In other words, mastering one or more attributes increases the mastery probability of the last attribute. Although an attribute map is not necessary to construct a

good assessment, a series of well-delineated sets of attributes and a set of well-developed items written to the attributes are essential. Ideally, the data out of the test is so well behaved and measured that the attributes are good enough to detect a map that could possibly be used for learning or instructional purposes.

To this purpose, this study examined the possible structural models of the attributes, using both real data and simulated data in the framework of Loglinear Cognitive Diagnosis Models (LCDMs) and Bayesian Networks. Additionally, this research explored methodology for possible attribute map structures that maximizes the impact of the map structure to the implementation and development of a diagnostic assessment. This study focused on the attribute structure itself, path links, to be exact. The goal is to make inferences from the hypothetical attribute map to a future learning map-based assessment construction for diagnosis purposes both in the area of instruction and learning.

The simulation study was constructed based on the structure of the real data. The steps in the simulation study were as follows: (1) generate a person profile of attribute mastery; (2) generate item parameters; (3) estimate model parameters and person profiles; (4) replicate steps one to three 100 times; and (5) evaluate model fit and parameter estimates by condition level. Three major factors were manipulated in the stimulation study: (1) sample size; (2) test length; and (3) attribute structure.

Both the parameter recovery and accuracy of student mastery classification were examined. Several indices for the accuracy and precision of the simulation study were provided. The indices of parameter recovery were as follows: (1) the bias; (2) the root mean squared error (RMSE); and (3) the 95% credible intervals of parameter estimates. The precision of student

mastery classification was examined by comparing student classification rate from the estimates with it from the true values.

This study addressed the following research questions:

1. How would a model with an over-specified set of pathways between attributes influence student mastery classification, item parameters, and map parameters?
2. How would a model with an under-specified set of pathways influence student mastery classification, item parameters, and map parameters?
3. Do estimated parameters in a misspecified model provide clues as to its misspecification?
4. What is the effect of the sample size and test length across models?

CHAPTER TWO

Literature Review

Diagnostic Classification Models (DCMs)

With the growing need for providing more information on test outcomes to test takers and pinpointing areas of strengths and weaknesses (Huff & Goodman, 2007; Trout & Hyde, 2006), diagnostic classification models (DCMs) have emerged as one technique that can be used to provide specific feedback on both the test taker's ability and areas for improvement. As the name suggests, diagnostic modeling provides the unique ability to "diagnose" or identify examinee's strengths and weaknesses with regard to the specific cognitive processes underlying performance on an assessment (Gierl, 2007; Yang & Embretson, 2007). The likelihood of mastery is represented by the probability of having mastered particular skills or attributes, with values closer to 1.0 indicating greater likelihood of skill mastery, and values approaching zero indicating likelihood of skill non-mastery.

DCMs are different from many common psychometric models (e.g., IRT) in that they contain categorical, rather than continuous, latent variables. Because of the categorization of their measured latent variables, DCMs yield respondent classifications with a high classification reliability for a small set of items. DCMs focus on why a respondent is not performing well instead of only focusing on who is performing well. The models define the chances of a correct response based on the respondent's attribute profile. The models also predict how respondents will answer each item. In the end, classification, rather than a single score, would be statistically inferred from DCMs. The result is a profile that indicates whether each respondent meets the criteria for diagnosis based on the attributes.

Elements of Diagnostic Classification Models

In DCMs, latent variables are often called *Attributes,* denoted as a $(a = 1, 2, \dots, A)$. Attributes are categorical latent variables representing the diagnostic statuses of a person. By taking a test, each respondent, denoted by the subscript *r*, will have multiple attributes measured by the items of the test, and these attributes altogether form an *Attribute profile* for that particular respondent, $a_r = [a_{r1}, a_{r2}, \dots, a_{rA}], a_{ra} \in \{0, 1\}$. The possible permutations of all attributes, (if all attribute permutations are possible) can be represented with $c = 2^A$.

To signify the relationship between items and the attributes needed for responses, an item-by-attribute Q-matrix is constructed. Q-matrix shows which item measures each attribute. The Q-matrix has the items in the rows and the attributes in the columns. An entry of "1" indicates that an attribute is measured by the corresponding item and vice versa. For example, suppose we have three math items, each measuring a few attributes that may or may not be shared by one another: addition, subtraction, multiplication, and division.

Item 1: 4+2 – 3

Item 2: 3 × 5

Item 3: (6/2) × 3

In item 1, the addition and subtraction attributes are measured, multiplication is measured by item 2, and division and multiplication are measured by item 3. Any attribute that has been measured by an item is denoted as "1" in an attribute pattern file versus non-measured as "0". As shown in Table 1, the Q-matrix row for each item, respectively, is: item one [1 1 0 0], item two [0 0 1 0], and item three is [0 0 1 1].

Table 2.1

An example of Q-matrix

	Addition	Subtraction	Multiplication	Division
4+2 – 3	1	1	0	0
3 × 5	0	0	1	0
(6/2) × 3	0	0	1	1

In general, DCMs are useful in terms of diagnosing students' abilities that are measured in a test, and thereby form a profile representing skills required for mastering each item in the test. Nevertheless, while DCMs conceptualize the attributes and diagnose mastery patterns with regard to these attributes, further modeling is still needed to estimate the probability of a person answering an item correctly. Loglinear Cognitive Diagnosis Model (LCDM) is the proper tool for modeling the probability of a correct response to an item.

Loglinear Cognitive Diagnosis Model (LCDM)

The Loglinear Cognitive Diagnosis Model (LCDM), based on the General Diagnostic Model (von Davier, 2005), models the probability of a correct response to an item as a function of the latent attributes of a respondent. In the model, item parameters consist of a set of main effects and interactions, linking attributes to item responses, providing a function of the probability of a correct response conditional on a respondent's attribute profile. In this sense, the LCDM uses an Analysis of Variance (ANOVA)-like approach to map latent attributes onto item responses, which predicts the dependent variable with observed factors as predictors. Following the ANOVA analogy, the LCDM treats the attributes as crossed experimental factors, which assumes all combinations of the attributes are a possible knowledge state of the examinee. The LCDM equation accepts any possible combinations of attributes measured by an item.

To illustrate how LCDM works, take our previous math item as an example: 4+2 – 3. The item measures two attributes, addition (attribute 1: α_{r1}) and subtraction (attribute 2: α_{r2}). The log-odds for respondent *r* mastering this item *i* is:

$$Logit(Y_{ri} = 1 \mid \alpha_r) = \lambda_{i,0} + \lambda_{i,1,(1)}\alpha_{r1} + \lambda_{i,1,(2)}\alpha_{r2} + \lambda_{i,2,(1,2)}\alpha_{r1}\alpha_{r2}$$

In the formula, the intercept $\lambda_{i,0}$ refers to the logit (log-odds) for non-masters of addition and subtraction. In other words, it refers to a respondent who has not mastered either of the attributes ($\alpha_{r1} = 0$ and $\alpha_{r2} = 0$). The term $\lambda_{i,1,(1)}$ is the conditional main effect for attribute 1, addition, indicating the increase in the logit of a correct response to the item for mastering addition for respondents who have not mastered subtraction. Likewise, $\lambda_{i,1,(2)}$ denotes the conditional main effect for attribute 2, subtraction, suggesting the increase in logit of a correct response to the item for mastering subtraction for respondents who have not mastered addition. The term $\lambda_{i,2,(1,2)}$ denotes to the 2-way interaction between the two attributes, addition and subtraction, representing the change in the logit for mastering both addition and subtraction.

There are other models subsumed by the LCDM that are commonly used for diagnosis. If the interaction is positive, we call that as an over-additive logit model, which includes conjunctive models and the DINA model (deterministic input noisy and-gate; Haertel, 1989). If the interaction is negative, we call that an under-additive logit model, which includes disjunctive models and DINO model (deterministic input noisy or-gate; Templin & Henson, 2006). Extreme interactions are unlikely in practice, whereas positive interactions with positive main effects are very likely. There are other models associated with LCDM besides the two mentioned above, NIDA model (noisy input deterministic and-gate; Maris, 1995), NIDO model (noisy input

deterministic or-gate; Templin & Henson, 2006), RUM (reparametrized unified model; Hartz, 2002), and the compensatory RUM (Hartz, 2002).

Comparison and Contrast of DCMs and IRT

In large-scale assessment, many useful statistical models have emerged for measuring latent variables from students. Item Response Theory (IRT) models are well known for their ability to model item responses as a function of both item and person characteristics (Lord, 2012). IRT has been widely used to predict a student's performance on an item using characteristics of the item (e.g., discrimination or difficulty parameters) and the abilities presumed to underlie performance on the test. However, the purpose of using the statistical models may vary from providing general information based on a single score or making diagnoses and classifications of students. On the contrary, diagnostic classification models (DCMs) provide students with profiles of mastery or non-mastery for a set of particular skills, in order to provide further information to learning and instruction.

The most salient distinction between IRT and DCMs is the nature of the latent variables, continuous and categorical respectively. IRT aims to place students on a continuous scale, whereas DCMs classify students into two categories (mastery or non-mastery) per attribute. In light of this difference, the reliability of IRT and DCM also differs. Reliability, by definition, refers to the consistency and repeatability of a test score. In other words, how much error is included in the test score. In IRT, reliability is usually mentioned as conditional reliability or conditional standard error of measurement. It is contingent upon test information, which is closely related to the number of items in a test. In DCMs, however, the latent variable is categorical, and the main purpose is classification rather than test scores.

Templin and Bradshaw (2013) demonstrated that DCMs provide a higher level of reliability for their estimates than comparable Item Response Theory (IRT) or Classic Test Theory (CTT) models. The statistical reason is that DCMs feature categorical variables, such that student estimates are reported in the form of Bernoulli variables (0 = non-mastery or 1=mastery). On the other hand, if categories are to be classified using IRT on a continuous scale, raters have to set standards (cut scores) based on the scores on the continuous scale. As a result, with the same number of items, DCMs manage to provide higher reliability than IRT. That also means that DCM are able to provide information for students with less data. Such characteristics allow DCMs to potentially change how large scale testing is conducted, which in turn reduces test development costs and student test taking errors.

Application of LCDM in Bayesian Networks

A Bayesian network is a directed acyclic graph (DAG) network, where the arrows represent hypothesized causal influences and represent induced and non-transitive dependencies by presenting useful independencies in the network (Pearl, 1998). The modeling of Bayesian networks is based on statistical probability theory. The notion of "information relevance" is represented by the conditional independence of the observed data, which offers an intuition about how dependencies should change in response to a given state of knowledge. In other words, a Bayesian network is a graphical model that visually represents the conditional independence relationships over the random variables in a joint distribution (Heckerman, 1995).

In the context of a structured attribute map, one skill state can be viewed as the precursor of mastering or proceeding to the next skill state. In addition, if all the skills and paths connect together, the map can be represented as a Bayesian Network. A snapshot of an attribute map is presented in Figure 2.1.

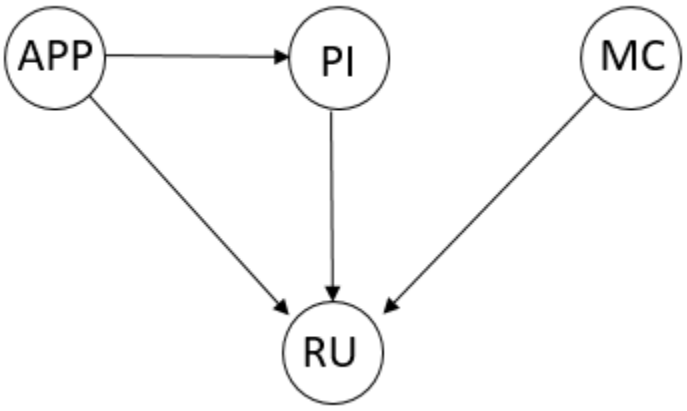

Figure 2.1. An example of the attribute map based on the four attributes.

When we look closer, as shown in Figure 2.1, we will see that each circle represents an attribute, which is a latent variable in DCMs, and the arrows represent the causal hypotheses between two attributes. Thereby, one subsequent attribute (also known as a child attribute) can be predicted or caused by one or more than one attribute; the precursor attribute (also known as a parent attribute) *PI* is connected to other attributes, and thus can be a subsequent attribute to other attributes as well in that context.

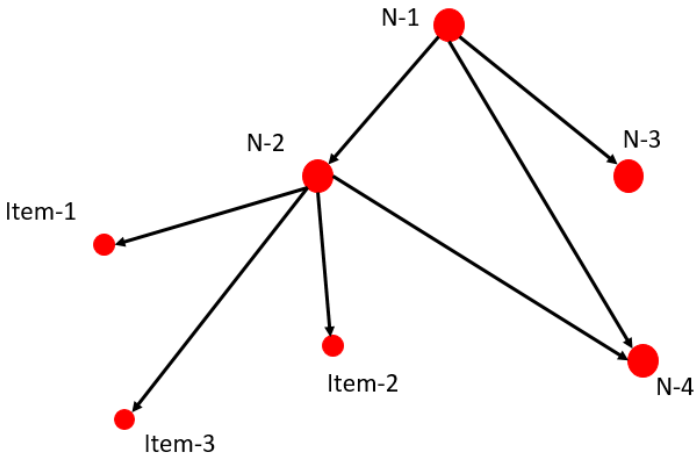

Figure 2.2. Closer look of attributes network

To be more specific, given the causal hypotheses between a parent attribute and the child attribute, we could assume the probability of mastering the child attribute should increase given the mastery of the parent attributes, as ruled by conditional probability theory. The caveat is that if the probability of the child attribute is high even without the predicting effect from the mastery of the parent attribute, the connection between these particular parent attribute and child attribute is implausible. Take attribute N-4 for example: attribute N-4 is a child attribute of two attributes, N-1 and N-2 as shown in Figure 2.2. Therefore, we expect that the probability of mastery of attribute N-4 is less than .5.

The conditional probability of child attribute mastery depending on the parent attribute is presented in Table 2.2 as an illustration. The probability of mastery of attribute N-4 is greater than 0.5, even with non-mastery of the parent attribute N-2. Therefore, the connection between these two attributes is questionable, in that the causal inference is not sustained. In other words, mastery of attribute N-2 does not necessarily have a causal effect on the likelihood of mastering N-4.

Table 2.2.

Example of conditional attribute probability

	N-4	
N-2	Master	Non-Master
Non-Master	.96	.04
Master	.99	.01

The same principle applies to item level estimates. The probability of mastering an item is conditioned on the probability of mastering the attribute the item is measuring. Take item 1 and item 2 as an example, there should be a higher probability the student will answer these items correctly if given mastery of attribute N-2. If this hypothesis does not hold in any items,

that is an indication that the item needs to be revisited by content experts, in that the item might not be appropriately measuring this attribute.

To summarize, once statistical evidence concerning recommendations for revisions to the map structure and items is compiled, decisions can be made regarding revisions to the attributes and their connections, as well as to the items. This study examined the original ordering of attributes and connections and went further to explore the possible attribute size and connections in the map structure, aiming to inform future research and operational work in assessment.

CHAPTER THREE

Methodology

The purpose of this study is to detect a possible map structure formed by the four attributes on fractional reasoning: Referent Units (RU), Partitioning and Iterating (PI), Appropriateness (APP), and Multiplicative Comparison (MC). More specifically, this research examined the specifications and connections of attributes and the possible pathways the attribute map should be using for optimal outcomes. The four attributes on fractional reasoning were the basis of constructing the models in simulation study. Empirical research has shown that sample size, test length, and model specification can affect the parameter and classification recovery of cognitive models (e.g., Choi, Templin, & Cohen, 2010; Kuniana-Jabenicht, 2012). Therefore, the simulation study at this point was conducted in varying conditions (sample size, test length, and model specification) to explore more options in attribute structure. In terms of statistical models, Loglinear Cognitive Diagnosis Models (LCDM) and Bayesian Networks were employed in both real data analysis and simulation study. The three factors involved in the data generation and estimation are listed below:

1. Sample size. Sample size was manipulated to assess the effect of varying sample size on the accuracy of the estimates of parameters and student classification. The sample size factor contains three levels: a large sample size, $N = 1000$ (based on real data), a medium sample size, $N = 500$, and a small sample size, $N = 100$.
2. Test length. Test length was varied on the attribute level because there was a limited number of items on certain attributes in the real data (i.e., there was just one item measuring attribute *MC* solely in the real data). Varying the test length will explore the effect of test length on parameter estimates accuracy and student classification in this study. Therefore, test length here refers to the number of items included within a

single attribute. In the real data, within an attribute, there were at least five items, and 27 items in total (see Figure 3.1). In light of the test length in the real data, three different test lengths were explored: n = 5 (small), n = 10 (medium), n=15 (large).

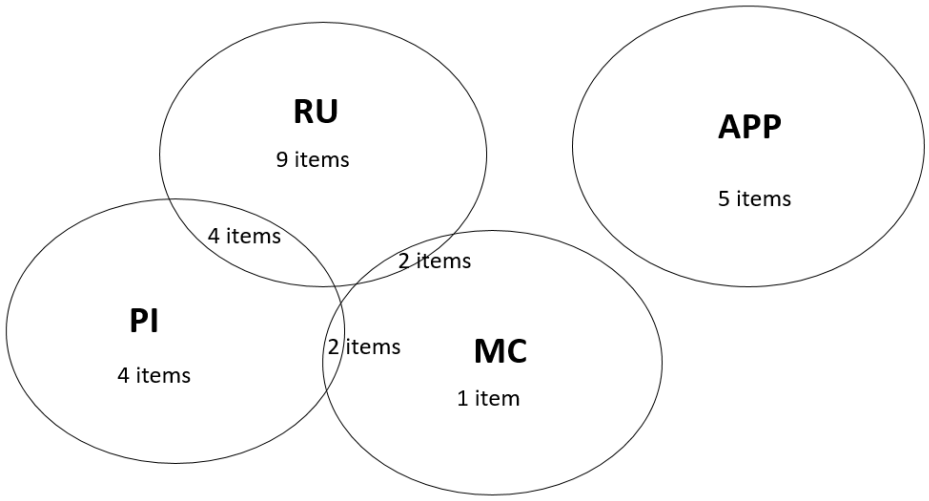

Figure 3.1. Venn diagram of item allocation among four attributes in real data

3. Structural models. Multiple structural models were listed here based on the possible specification of pathways in an attribute diagram. As mentioned above, the attribute structure forms a learning map type diagram that guides the test development and thus determines student mastery classification; how the pathways specified in the diagram extends an impact on this matter. Thereby, this study explored additional three possible ways (over-specified pathways in an attribute structure, under-specified pathways in an attribute structure, and linear structure) an attribute diagram should be structured to assess the effect of pathway specification on parameter estimation and student mastery classification. Each data set was generated using the hypothetically correct model varying sample size and test length and was then estimated in the

following statistical models. Models used for data estimation in simulation study are listed below.

a) Hypothetically correct model informed by real data. According to the real data, attributes *APP*, *PI*, and *MC* were independent attributes serving as parent attributes to attribute *RU*. Therefore, to be congruent with the real data, and to explore more possible map structures based on the real data, the model informed by the real data is used as a correct model.

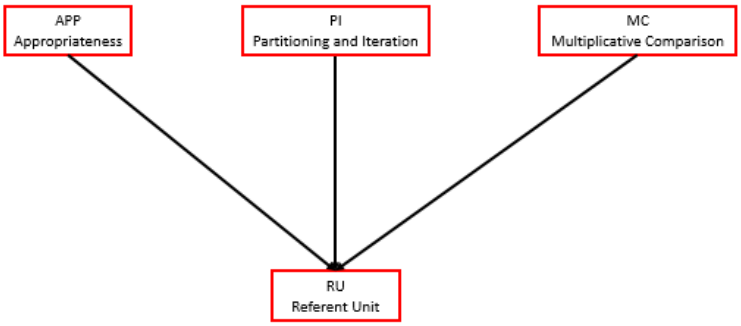

Figure 3.2. Hypothetically correct model

b) Model with over-specified pathways. With an "over-specified pathways" model, one more path link was randomly added into the hypothetically correct model, indicated in blue. Therefore, instead of having all three attributes independent, attribute *PI* was dependent upon attribute *APP*.

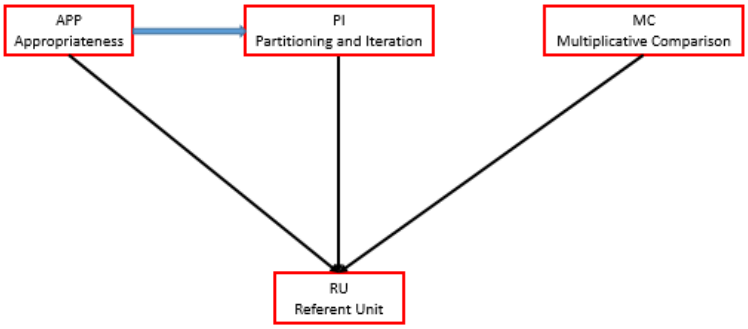

Figure 3.3. Model with over-specified pathways.

c) Model with under-specified pathways. One of the pathways was randomly deleted to qualify as under specified from our correct model, indicated by a blue dotted line. By deleting one pathway from the hypothetically correct model, attribute *MC* turned to be isolated from the model and not connected to any other attributes; this also provided information about the specification of the attribute grain size to the student mastery classification. The detailed results are discussed in subsequent chapters.

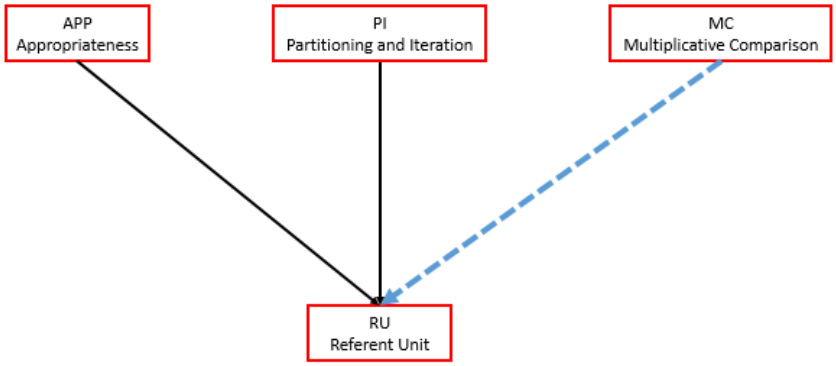

Figure 3.4. Model with under specified pathways.

d) Misspecified model. A linear model is chosen to be an example of a misspecified model in this case, in that the real data does not indicate a linear attribute structure, and linear model is the most commonly seen and used model in the field of assessment and is not the right model in this case due to the property of the real data. The real data was built based on a multidimensional test framework, and therefore a linear model was not feasible, but was included in simulation study for cross validating the attribute structure.

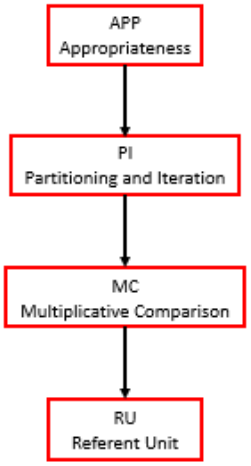

Figure 3.5. Linear model.

In general, the different levels of the two controlled factors yielded $3 \times 3 = 9$ data generation conditions. Each data set was estimated by the four models: the correct model, the over-specified model, the under-specified model, and the linear model, resulting in a total of 36 conditions explored in this simulation study. Table 3.1 summarized the manipulated factors with their levels in the simulation study.

Table 3.1

The manipulated factors in the simulation study

	Factors	Factor level	Number of levels
1	Sample size	Large: 1000 Medium: 500 Small: 100	3
2	Test length	Long: 15 Medium: 10 Short: 5	3
3	Attribute structure	Correct model Over-specified model Under-specified model Linear model	4

For the simulation study, the uniform distribution for attribute profiles in data generation is one that has been used in many of the current simulation studies in DCMs (e.g., de la Torre, 2009). The uniform distribution indicates that every profile has equal probability—mastery or non-mastery. The range of the uniform distribution for attribute intercept was chosen [−1, 1] to give symmetry around a log-odds of zero (p = 0.5) *for* masters of few attributes and masters of most attributes (Templin & Bradshaw, 2014). Same principle applies to item intercept. In terms of main effect and interaction's uniform distribution range of [0, 2], the range was picked to reflect the probability of mastery of 0.5 and beyond but does not exceed the probability of 1 when counting in probability of guessing (intercept). The steps in the simulation study are listed below:

1. Generate person profile. Person profile of attribute mastery (a vector of binary numbers) for a simulated student r, α_r, was generated from a series of conditional univariate Bernoulli distributions $\alpha_{ra} \sim B\left(p_{\alpha_{ra}}\right)$, where $p_{\alpha_{ra}}$ is the probability student r is a master of attribute a ($a = 1, 2, \ldots, A$). The probability $p_{\alpha_{ra}}$ is conditional on the value of a "prerequisite" attribute a' defined by:

$$p_{\alpha_{ra}} = \frac{\exp\left(\gamma_{a,0} + \gamma_{a,1,(a')}\alpha_{ra'}\right)}{1 + \exp\left(\gamma_{a,0} + \gamma_{a,1,(a')}\alpha_{ra'}\right)}$$

a. When the attribute does not have a predictor, the probability is no longer conditional on attribute a' and $\gamma_{a,1,(a')} = 0$. Each attribute intercept, $\gamma_{a,0}$, was randomly drawn from a uniform distribution with a range of [−1, 1].
b. Each attribute main effect, $\gamma_{a,1,(a')}$, was randomly drawn from a uniform distribution with a range of [0, 2].
c. When there are interaction terms involved between attributes, for instance, in the hypothetically correct model, attribute *RU* was dependent upon the any of the combination of the remaining three attributes – *APP*, *PI*, and *MC*, the interaction effect was also randomly drawn from a uniform distribution with a range of [0, 2].

Take the hypothetically correct model for an example (see Figure 3.2), attribute *APP* is not conditioned on any other attributes, and therefore the probability of *APP* can be defined by:

$$p_{\alpha_{ra}} = \frac{\exp\left(\gamma_{a,0}\right)}{1 + \exp\left(\gamma_{a,0}\right)}$$

For attribute *RU*, who has three parent attributes, and thus includes three main effects, three two-way interactions and one three-way interaction. The logit can be expressed by:

$$Logit\left(Y_{ra} = 1 \mid \alpha_{ra}\right) = \gamma_{a,0} + \gamma_{a,1,(a1')}\alpha_{ra1'} + \gamma_{a,1,(a2')}\alpha_{ra2'} + \gamma_{a,1,(a3')}\alpha_{ra3'}$$
$$+\gamma_{a,2,(a1',a2')}\alpha_{ra1'}\alpha_{ra2'} + \gamma_{a,2,(a1',a3')}\alpha_{ra1'}\alpha_{ra3'} + \gamma_{a,2,(a2',a3')}\alpha_{ra2'}\alpha_{ra3'} +$$
$$\gamma_{a,3,(a1',a2',a3')}\alpha_{ra1'}\alpha_{ra2'}\alpha_{ra3'}$$

And the probability of *RU* can be defined by:

$$p_{\alpha_{ra}} = \frac{\exp\left(Logit\left(Y_{ra} = 1 \mid \alpha_{ra}\right)\right)}{1 + \exp\left(Logit\left(Y_{ra} = 1 \mid \alpha_{ra}\right)\right)}$$

2. Generate item parameters. Item parameters were generated following the same principle. Conditional on a person's mastery profile, α_r, generated from step one, item responses were generated from a series of conditional univariate Bernoulli distributions, where, for an item i, measuring only one attribute, $Y_{ri} \sim B\left(p_{Y_{ri}}\right)$, with:

$$p_{Y_{ru}} = \frac{\exp\left(\lambda_{i,o} + \lambda_{i,1,(a)}\alpha_{ra}\right)}{1 + \exp\left(\lambda_{i,o} + \lambda_{i,1,(a)}\alpha_{ra}\right)}$$

   a. Each intercept, $\lambda_{i,0}$, was randomly drawn from a uniform distribution with a range of [-1, 1].
   b. Each conditional main effect, $\lambda_{i,1,(a_{1,\ldots})}$, was randomly drawn from a uniform distribution with a range of [0, 2].
   c. When there is interaction terms involved, the interaction effect was also randomly drawn from a uniform distribution with a range of [0, 2].

If an item i¸ measures two attributes, α_{r1}, and α_{r2}, the main effect for both attributes along with the interaction between the two attributes will be involved in calculating the probability:

$$p_{Y_{ru}} = \frac{exp\left(\lambda_{i,o} + \lambda_{i,1,(a1)}\alpha_{ra1} + \lambda_{i,1,(a2)}\alpha_{ra2} + \lambda_{i,2,(a1,a2)}\alpha_{ra1}\alpha_{ra2}\right)}{1 + exp\left(\lambda_{i,o} + \lambda_{i,1,(a1)}\alpha_{ra1} + \lambda_{i,1,(a2)}\alpha_{ra2} + \lambda_{i,2,(a1,a2)}\alpha_{ra1}\alpha_{ra2}\right)}$$

3. Estimate model parameters and person profiles. The data set was analyzed using both LCDM and Bayesian Networks.
4. Replication. Step one to step three were repeated 100 times and computed the model fit, parameter recovery, and student mastery classification for each data set.

Data Analysis

The Bayesian Networks estimated map parameters and student mastery status simultaneously. The sections below describe the estimation procedures using Bayesian Networks. In the MCMC algorithm, the priors for the attribute and item parameters follow a normal distribution because the parameters are on the logit scale. The priors for attribute parameters (intercept, main effect, and interaction) and item intercept were drawn from a normal distribution with mean zero and precision of 0.1, which is not informative. The priors for item main effects and item interactions were drawn from truncated normal distribution with mean zero and precision of 0.1, because item main effects and interactions range from zero to positive infinity. Two Markov chains were drawn in the estimation. One thousand iterations of burn-in were used for each run. A total of 5000 iterations were run to obtain samples for posterior inferences. No thinning was done during the MCMC process, and thus all chains were retained for inference making (as recommended by Gelman et al., 2004). As for checking the convergence of the estimation, the percentage of R-hat values greater than 1.2 were calculated. Gelman et al. recommended R-hat values less than 1.2 as good evidence of convergence (Brooks & Gelman, 1998), or less than 1.1 (default value of JAGS).

Parameter estimates were computed using the MCMC algorithm. The concept of MCMC estimation is to construct a set of random draws from the posterior distribution of each parameter being estimated, and a stationary distribution of the random draws is desired. Using the draws from the posterior distribution of parameters, point estimates can then be calculated. Before the chains converge to a stationary distribution, a "burn-in" in the MCMC process discards the first thousand draws because early on the random draws might wander around low-density parameter space. Once the chains converge, the remaining draws are kept for making inferences.

Evaluation criteria

This section discusses the criteria for (a) model fit indices, (b) student mastery classification accuracy, and (c) map and item parameter recovery. The criteria were used for both the simulation study and the real data analysis.

The DIC (deviance information criterion; Spiegelhalter et al., 2002) was computed for goodness-of-model-fit index, and was provided by JAGS (Just Another Gibbs Sampler, Plummer, 2012). Student mastery classification rate was calculated across the conditions across the models. The classification rate was then used to compare with the classification rate of true values to check the accuracy of student mastery classification. In the meantime, bias and Root Mean Squared Error (RMSE) were computed between the person profile estimates and the true values for the attributes.

The indices and figures of the precision of map and item parameter estimates were reported for each condition. And these indices were: a) average bias of the map and item parameters; b) average RMSE of the parameters; and c) average 2.5% and 97.5% thresholds of the credible intervals. Bias was calculated using the formula: $B(\zeta) = \hat{\zeta} - \zeta$, with $\hat{\zeta}$ denoting the mean of estimates $\hat{\zeta} = \frac{1}{100}\sum \widehat{\zeta_\iota}$, and ζ the mean of true value of the parameter $\zeta = \frac{1}{100}\sum \zeta_i$, where

i refers to the sequence of data set. The root mean squared error (RMSE) was defined with the formula: $RMSE(\hat{\zeta}) = \sqrt{\frac{1}{100}\sum(\hat{\zeta} - \zeta)^2}$, which represents the degree of deviation of estimates from the true values of parameters. The credible intervals were provided by JAGS, and were computed after an average over the replications was taken. In addition to bias and RMSEs to examine parameter recovery, one-way Analysis of Variance (ANOVA) was employed in SPSS to test differences in recovery as a function of the conditions. Eta-squared (η^2) was calculated to quantify how much each factor – sample size, test length, and structure specification – mattered empirically.

The computer language R (R Development Core Team, 2013), JAGS (Just Another Gibbs Sampler, Plummer, 2012), R2jags package (Su & Yajima, 2015), and rjags package (Plummer, 2013) were used for data generation and data analysis of LCDM and Bayesian Networks.

CHAPTER FOUR

Simulation Results

In the simulation study, three aspects were manipulated: (1) the sample size; (2) the test length; (3) and the models. Thirty-six data sets were generated combining these factors. Each data set was then analyzed with LCDM integrated Bayesian networks, and replicated 100 times. Bias and RMSEs were calculated for each condition and all the results were tabled accordingly.

The Measures of Convergence for the MCMC Algorithm in the Bayesian Network

The R-hat indices of convergence for the MCMC algorithm are summarized and presented in Table 4.1. Table 4.1 gives the percentage of R-hat values that are greater than 1.2 for each set of parameters to indicate a bad convergence, and the maximum R-hat in the parameters set from all data sets. The convergence was examined at three levels: (1) overall model fit–the deviance; (2) evidence model fit–item parameters; and (3) proficiency model fit – attribute parameters and person mastery classification.

At the model level, as shown in Table 4.1, the R-hat values of the deviance for all the conditions across all four models were close to 1 and always less than 1.2, with the largest value of 1.17. Hence, the estimations at the model level converged well in terms of R-hat values of the deviances. In addition, because the statistical models for data generation and estimation were the same, the estimates for the evidence model and proficiency model showed a good convergence as well. More specifically, the R-hat values for attribute parameters and item parameters were consistently below 1.2, with the highest values of 1.13 and 1.11, respectively.

In addition, holding sample size constant, as the test length increases, the convergence does not improve. As the test length increases, the proficiency model converges worse, if model

conditions is held constant. Overall, all parameters converged well across conditions across models as anticipated.

Table 4.1

The percentage of R-hat values greater than 1.2 and the maximum R-hat for each parameter set

Condition			Deviance		Lambda0		Lambda1		Lambda11		Gamma0		Gamma1		Gamma11	
Model	Sample Size	Test Length	Percent	Max	Percent	Max	Percent	Max	Percent	Max	Percent	Max	Percent	Max	Percent	Max
Correct Model	100	28	0	1.04	0	1.02	0	1.01	0	1.01	0	1.03	0	1.04	0	1.02
		48	0	1.06	0	1.01	0	1.01	0	1.01	0	1.06	0	1.05	0	1.03
		68	0	1.05	0	1.02	0	1.01	0	1.01	0	1.05	0	1.06	0	1.02
	500	28	0	1.06	0	1.06	0	1.01	0	1.01	0	1.06	0	1.06	0	1.06
		48	0	1.06	0	1.04	0	1.02	0	1.01	0	1.08	0	1.07	0	1.05
		68	0	1.08	0	1.04	0	1.02	0	1.01	0	1.05	0	1.08	0	1.05
	1000	28	0	1.10	0	1.09	0	1.04	0	1.02	0	1.08	0	1.10	0	1.07
		48	0	1.09	0	1.07	0	1.02	0	1.02	0	1.07	0	1.09	0	1.09
		68	0	1.09	0	1.05	0	1.02	0	1.02	0	1.08	0	1.09	0	1.09
Over-specified Model	100	28	0	1.04	0	1.02	0	1.01	0	1.01	0	1.03	0	1.04	0	1.03
		48	0	1.06	0	1.03	0	1.01	0	1.01	0	1.06	0	1.05	0	1.03
		68	0	1.06	0	1.02	0	1.01	0	1.01	0	1.06	0	1.04	0	1.02
	500	28	0	1.08	0	1.04	0	1.03	0	1.02	0	1.05	0	1.08	0	1.05
		48	0	1.07	0	1.06	0	1.02	0	1.01	0	1.06	0	1.07	0	1.05
		68	0	1.07	0	1.04	0	1.01	0	1.01	0	1.06	0	1.07	0	1.06
	1000	28	0	1.17	0	1.13	0	1.04	0	1.03	0	1.11	0	1.17	0	1.08
		48	0	1.08	0	1.06	0	1.05	0	1.02	0	1.06	0	1.08	0	1.07
		68	0	1.10	0	1.04	0	1.02	0	1.01	0	1.09	0	1.10	0	1.08
Under-specified Model	100	28	0	1.03	0	1.02	0	1.01	0	1.01	0	1.03	0	1.03	0	1.03
		48	0	1.05	0	1.01	0	1.01	0	1.01	0	1.03	0	1.05	0	1.03
		68	0	1.04	0	1.02	0	1.01	0	1.01	0	1.03	0	1.04	0	1.02
	500	28	0	1.10	0	1.04	0	1.02	0	1.02	0	1.10	0	1.07	0	1.05
		48	0	1.09	0	1.04	0	1.02	0	1.01	0	1.05	0	1.09	0	1.08
		68	0	1.07	0	1.04	0	1.01	0	1.01	0	1.05	0	1.07	0	1.04
	1000	28	0	1.12	0	1.11	0	1.08	0	1.02	0	1.09	0	1.12	0	1.08
		48	0	1.09	0	1.06	0	1.05	0	1.02	0	1.08	0	1.09	0	1.09
		68	0	1.09	0	1.05	0	1.02	0	1.02	0	1.07	0	1.09	0	1.08
Linear Model	100	28	0	1.06	0	1.04	0	1.02			0	1.06	0	1.04	0	1.03
		48	0	1.04	0	1.02	0	1.01			0	1.03	0	1.04	0	1.03
		68	0	1.04	0	1.01	0	1.01			0	1.04	0	1.04	0	1.02
	500	28	0	1.11	0	1.11	0	1.03			0	1.04	0	1.08	0	1.05
		48	0	1.07	0	1.06	0	1.02			0	1.07	0	1.07	0	1.06
		68	0	1.07	0	1.03	0	1.02			0	1.05	0	1.07	0	1.05
	1000	28	0	1.10	0	1.09	0	1.04			0	1.10	0	1.10	0	1.07
		48	0	1.09	0	1.07	0	1.03			0	1.06	0	1.07	0	1.09
		68	0	1.08	0	1.06	0	1.03			0	1.06	0	1.07	0	1.08

Note. Lambda0=attribute intercept; Lambda1=attribute main effect; Lambda11=attribute interaction; Gamma0=item intercept; Gamma1=item main effect; Gamma11=item interaction.

Deviance

The goodness-of-fit of the models were compared using DIC, which is generated by JAGS. The smaller the DIC is, the better the model fits the data. Table 4.2 presents the means of DIC values for the models across conditions. Following the DIC rule of the smaller the better, under-specified model fits slightly better than the rest of the models, with the small sample size and short test length. However, as the condition goes more complex, the goodness-of-fit index gets penalized by generating bigger DICs. To this extent, the simplest structure – the linear structure – gets the smallest DIC when the sample size is large and the test length is long. More detailed model-data fit information is to be discussed in the following sections.

Table 4.2

The mean of DIC across conditions

Condition			
Model	Sample Size	Test Length	DIC
Correct Model	100	28	3285.93
		48	5722.23
		68	8153.72
	500	28	15940.58
		48	27777.00
		68	39648.52
	1000	28	31650.60
		48	55188.42
		68	78612.57
Over-specified Model	100	28	3294.08
		48	5739.96
		68	8207.55
	500	28	15921.04
		48	27760.70
		68	39811.33
	1000	28	31306.07
		48	55568.20
		68	78948.95
Under-specified Model	100	28	3236.66
		48	5736.45
		68	8110.90
	500	28	16043.10
		48	27995.25
		68	39758.02
	1000	28	31455.74
		48	55254.64
		68	78737.93
Linear Model	100	28	3250.70
		48	5717.92
		68	8182.68
	500	28	16075.30
		48	27819.38
		68	39694.31
	1000	28	31436.11
		48	55328.10
		68	78645.07

Item Level Results

The item level results include the estimation of the item intercept, main effect, and interactions between items and corresponding attributes. For each condition, (1) the empirical bias $B(\hat{\zeta})$, (2) the root mean squared error (RMSE), (3) and the percentage of 95% credible intervals, were computed and reported in Table 4.3. Table 4.3 presents the means of these indices for the purpose of evaluating the recovery and precision of item parameters.

Results for Item Intercept Parameters

For item intercept parameters across all four models, the bias values are small and positively biased, meaning item intercept parameters are more likely to be slightly overestimated varying conditions and models. Given the same sample size, the bias tends to decrease slightly as the test length increases across models. Also, controlling the test length, the bias values decrease as the sample size increases across models. Likewise, the root means squared error (RMSE) follows a similar pattern. Holding the sample size in constant, the RMSEs decrease slightly as the test length increases, and vice versa.

It is interesting to note that in the correct model, the condition where the sample size is large and the test length is medium to large owns relatively smaller bias. The same finding applies to the over-specified model, and we suspect that this condition might be the most accurate one for estimating item intercept for pathways specified in the correct model and in the over-specified model. On the other hand, the changes in bias and RMSE in the under-specified model and the linear model showed a more consistent pattern. More specifically, as the sample size or test length goes up, the bias and RMSE tend to continually decrease. Moreover, given the same test length, the under-specified model shows slightly smaller bias, suggesting more accuracy in item intercept estimation among all four models. In short, the most accurate item

intercept estimation is when the sample size is large, and the test length is large, ideally in an under-specified model with a 95% credible interval of 45.8%. The 95% credible intervals ought to endorse for the parameter estimation, yet in item intercept estimation, the credible interval covers between 44.9% to 84.9% percent of the true values of intercept, across models and conditions. This makes us believe that there is no reliable evidence that any of the pathways specified in the models provide a highly accurate item intercept estimation in various models and conditions.

Results for Item Main Effect Parameters

In general, item main effect bias values across all four models and conditions were negative, indicating that the item main effect parameters were likely underestimated regardless of the variations in sample size, test length, and estimation models. When holding test length in constant, bias values decrease as the sample size increases the same as RMSE values across all four models. According to the bias values, the best model and condition for the most accurate item main effect estimation is the correct model with large sample size and long test length, followed by the same conditions in the under-specified model, the linear model, and the over-specified model.

Following the same pattern found in item main effect bias, RMSEs decrease as the sample size increases, holding test length constant. If holding sample size constant, RMSE values decrease as test length increases. The 95% credible intervals covers 64.1% to 80.5% of the true main effect values, suggesting a fairly good accuracy in estimating main effect. Overall, none of the models seem to perform better than others in estimating item main effect parameters as far as accuracy is concerned. The item main effect parameter estimation seem to have more preference in sample size and test length factors, rather than pathway variations in the models.

Results for Item Interaction Parameters

Similar to the item main effect parameters, item interaction parameters tended to be negatively biased and underestimated across models and conditions. In general, the bias decreases as the sample size increases, holding the test length in constant. It is interesting to note that for the correct model and the under-specified mode, the condition with large sample size and medium to long test length yielded smaller bias values, indicating the best accuracy in item interaction parameter estimation. On the other hand, the over-specified model and the linear model showed a more consistent decrease in bias values as the sample size or test length increases. Therefore, we suspect that the medium test length and large sample size have the most effect on item interaction estimation, especially in the correct model and the under-specified model. As for RMSE, the values dropped when the sample size or test length increases, holding the other one constant.

The 95% credible intervals cover from 75.7% to 88.1% of the true values. Within the same model, as the sample size goes up, the 95% CI coverage goes up as well, holding test length in constant. In contrast, as the test length goes up, the 95% CI coverage goes down, controlling for the sample size. This makes us suspect that item interaction estimation accuracy is more dependent upon the conditions – sample size and test length in our case, and the best condition for a more accurate item interaction estimation is large sample size along with small test length. In general, the best condition for item interaction estimation in the linear model is when the test length is long and the sample size is large, concerning the magnitude of the bias and RMSE. However, based on what we see in the 95% credible interval, the model probably does not have much effect on item interaction estimation.

Table 4.3

The mean of the bias, the RMSE, and the 95% creditable intervals for item intercept, main effect and interaction parameters

Condition			Intercept			Main Effect			Interaction		
Model	Sample Size	Test Length	Bias	RMSE	Percent	Bias	RMSE	Percent	Bias	RMSE	Percent
Correct Model	100	28	0.230	0.700	84.6	-0.634	0.855	80.5	-0.690	0.888	75.6
		48	0.231	0.727	84.4	-0.636	0.850	79.7	-0.646	0.848	77.2
		68	0.222	0.700	84.5	-0.626	0.852	79.6	-0.626	0.822	78.1
	500	28	0.228	0.728	63.6	-0.458	0.751	84.3	-0.495	0.720	87.3
		48	0.197	0.736	60.6	-0.411	0.724	80.1	-0.471	0.699	87.9
		68	0.172	0.712	57.3	-0.374	0.716	75.9	-0.457	0.682	87.1
	1000	28	0.141	0.728	53.8	-0.342	0.709	77.2	-0.423	0.672	88.0
		48	0.097	0.713	48.4	-0.284	0.686	69.5	-0.376	0.646	86.1
		68	0.117	0.718	45.2	-0.248	0.677	64.5	-0.399	0.646	85.3
Over-specified Model	100	28	0.269	0.738	84.9	-0.651	0.863	80.0	-0.702	0.885	75.8
		48	0.254	0.724	84.1	-0.630	0.857	79.2	-0.657	0.860	77.2
		68	0.227	0.703	84.6	-0.629	0.846	79.4	-0.653	0.856	78.3
	500	28	0.241	0.735	62.9	-0.463	0.754	83.9	-0.523	0.748	87.7
		48	0.192	0.727	59.6	-0.417	0.737	80.2	-0.460	0.697	87.2
		68	0.145	0.708	57.4	-0.382	0.714	76.2	-0.455	0.672	86.7
	1000	28	0.145	0.723	53.7	-0.345	0.709	77.3	-0.378	0.652	87.1
		48	0.141	0.719	48.7	-0.284	0.682	69.1	-0.386	0.638	85.8
		68	0.081	0.700	45.8	-0.254	0.687	64.1	-0.370	0.639	85.0
Under-specified Model	100	28	0.269	0.738	84.9	-0.635	0.856	80.9	-0.686	0.880	75.7
		48	0.254	0.724	84.1	-0.619	0.842	79.6	-0.656	0.856	76.3
		68	0.227	0.703	84.6	-0.624	0.849	80.0	-0.670	0.867	78.1
	500	28	0.241	0.735	62.9	-0.457	0.752	83.7	-0.506	0.733	86.6
		48	0.192	0.727	59.6	-0.407	0.728	79.7	-0.508	0.731	87.3
		68	0.145	0.708	57.4	-0.373	0.705	76.5	-0.464	0.703	87.2
	1000	28	0.145	0.723	53.7	-0.330	0.716	76.9	-0.420	0.667	88.0
		48	0.141	0.719	48.7	-0.277	0.685	69.6	-0.378	0.631	85.8
		68	0.081	0.700	45.8	-0.248	0.672	64.7	-0.385	0.630	85.0
Linear Model	100	28	0.301	0.717	85.0	-0.633	0.853	80.7	-0.691	0.883	75.9
		48	0.225	0.717	84.6	-0.630	0.852	79.9	-0.654	0.855	77.0
		68	0.234	0.726	84.4	-0.626	0.847	79.6	-0.667	0.863	77.3
	500	28	0.231	0.722	63.7	-0.470	0.761	84.2	-0.462	0.725	87.1
		48	0.178	0.728	60.1	-0.425	0.736	80.3	-0.447	0.694	87.8
		68	0.170	0.700	57.3	-0.380	0.710	76.5	-0.431	0.682	86.9
	1000	28	0.107	0.721	54.1	-0.327	0.708	77.2	-0.432	0.674	88.1
		48	0.141	0.724	48.3	-0.281	0.695	69.0	-0.379	0.639	86.0
		68	0.117	0.722	45.7	-0.250	0.677	64.8	-0.365	0.637	84.9

Note. RMSE=root mean squared error; Percent=percentage of 95% intervals, computed using the 2.5% and 97.5% thresholds of credible interval, covering the true values.

In this section pertaining to item parameters (item intercept, item main effect, item interaction) estimation results, all item parameters were estimated quite accurately using LCDM and Bayesian Networks. The item intercept parameters were positively biased, but own the least magnitude of bias and RMSEs, indicating item intercept parameters were more accurately estimated than main effect and interaction parameters across models. Due to the fact that some items measure more than one attribute, interaction terms were also estimated. All item interaction terms across models were negatively biased, though not with significant distance from the true values. In addition, the 95% credible intervals covered a majority of the true values for each set of item parameters: intercept, main effect, and interaction. Different models showed their own merits in estimating item parameters accurately; for instance, in terms of the magnitude of the bias across the four models, the over-specified model and the under-specified model estimated slightly more accurately in item intercept, the under-specified model outperformed the other three models in item main effect estimation, and the linear model was slightly more accurate in estimating item interaction parameters. Although there is no distinct evidence showing the difference of models in terms of item parameter estimation, underlying patterns concerning sample size and test length were explored in this section. In general, large sample size along with medium to long test length is preferred in estimating item parameters.

The findings from the previous sections pertaining to the impact of model specification, sample size, and test length on test item parameter recovery indicate that how a test was constructed based on a series of attributes structure does not have much effect on item level. The results were confirmed by the ANOVA results shown in Table 4.4. As shown in Table 4.4, sample size owns the biggest effect on item recovery, especially item main effect, followed by item interactions. Test length has medium effect on item main effect, and attribute structure

specification does not have much effect on item parameter recovery. The findings point out that in constructing an assessment, it is less important on how the map structure is in terms of skill connection and dependency, it is more important to ensure that items are well written and measuring what they are supposed to measure. More results about attribute parameter estimation and person mastery status will be discussed in the next section.

Table 4.4

Eta Squared for Item Recovery Parameters by Condition

		Sample Size	Test Length	Attribute Structure
Bias	Intercept	0.06	0.01	0
	Main Effect	0.82	0.26	0
	Interaction	0.28	0.01	0.01
RMSE	Intercept	0	0	0.0
	Main Effect	0.58	0.02	0.0
	Interaction	0.26	0.01	0.0

Note. Rule of thumb for η^2: 0.02=small, 0.13=medium, 0.26=large (Cohen, 1998; Miles & Shevlin, 2001)

Attribute Level Results

The attribute level results include the estimation of the attribute intercept, main effect, and interactions between attributes. For each of the results in each condition, (1) the empirical bias $B(\hat{\zeta})$, (2) the root mean squared error (RMSE), (3) and the percentage of 95% credible intervals, were computed and reported in Table 4.5. Table 4.5 presents the means of these indices, which were examined to evaluate the recovery and precision of attribute parameters.

Results for Attribute Intercept Parameter

First, for the hypothetically correct model, it is interesting to note that with small sample size, the bias goes up a bit as the test length increases. This pattern does not apply to the conditions where the sample size is medium or large. Also, when the sample size is small or

medium, the conditions with medium test length yields the least magnitude of bias and RMSE. Although if we hold medium or large sample size in constant, the bias decreases as the test length increases. The same principle applies to RMSEs. When the test length increases, there is also a significant decrease in RMSE values holding sample size in control. Besides, given the same test length, the RMSEs decrease as the sample size increases. In short, with medium to large sample size, longer test length yields smaller bias and RMSE, suggesting that the best condition to the correct model is when the sample size is medium to large and the test length is long.

Secondly, for the over-specified model, the small to medium sample size is not preferable because within the small sample size, the bias goes up as the test length increases and both the bias and RMSE values results from the medium sample size have the greatest magnitude. With that being said, the best condition for the over-specified model to estimate attribute intercept parameters is when the sample size is large, and the test length is medium. In other words, with a more complicated model, over-specified model in our case, a more complex condition is needed for the attribute intercept estimation to be more accurate.

Third, for the under-specified model, it is apparent that attribute intercept parameter estimates have the greatest magnitude among all models, indicating that the under-specified model does not perform as well as other models (except for the linear model, which was assumed as a wrong model in the multidimensional framework) in terms of attribute intercept estimation. When test length is short, the bias and RMSE increase as the sample size increases, which indicates that for the under-specified model, the short test length is not preferable in terms of attribute intercept estimation. When the test length is medium to long, the bias and RMSE values decrease as the sample size increases, which confirmed what was mentioned previously that a

more complex condition is needed when the model itself is not complex enough. As for the sample size effect, within medium to large sample size, as test length increases, bias and RMSE values decrease. Based on the smallest bias and RMSE, and the 95% credible interval, the condition of medium sample size and long test length is more preferable than other conditions for the under-specified model.

Lastly, the linear model showed the greatest magnitude of bias and RMSE in attribute intercept estimation as anticipated, meaning that the linear model is the least accurate model when it comes to estimating attribute intercept accurately. Although in the linear model, test length and sample size still showed their effect on parameter estimation accuracy. For instance, when the sample size is medium to large, the bias and RMSE decrease as the test length increases; when holding the test length in constant, the bias and RMSE values decrease as the sample size increases. In short, the condition where the sample size is large and the test length is medium yielded the best results for the linear model, although compared to other models, this linear model is not the best one. Overall, all intercept parameter estimation across models were estimated well, with one model performed better than the other, and with the 95% credible intervals covering up to 82.0% of the true values. Therefore, the attribute intercept parameters were reliably estimated across conditions.

Table 4.5

The mean of the bias, the RMSE, and the 95% creditable intervals for attribute intercept, main effect and interaction parameters

Condition			Intercept			Main Effect			Interaction		
Model	Sample Size	Test Length	Bias	RMSE	Percent	Bias	RMSE	Percent	Bias	RMSE	Percent
Correct Model	100	28	0.379	0.643	58.4	-0.787	0.939	49.0	-0.823	0.976	27.3
		48	0.231	0.554	68.3	-0.636	0.851	55.7	-0.646	0.988	31.9
		68	0.425	0.605	75.0	-0.697	0.855	61.7	-0.845	0.981	35.3
	500	28	0.366	0.587	73.4	-0.673	0.820	72.7	-0.744	0.908	49.5
		48	0.292	0.589	73.5	-0.609	0.765	85.1	-0.736	0.894	56.2
		68	0.295	0.608	60.6	-0.561	0.687	92.5	-0.682	0.840	62.3
	1000	28	0.357	0.630	64.7	-0.616	0.763	84.1	-0.724	0.900	59.4
		48	0.221	0.570	51.2	-0.493	0.633	99.7	-0.692	0.878	67.2
		68	0.138	0.532	40.7	-0.467	0.589	93.0	-0.675	0.846	72.1
Over-specified Model	100	28	0.288	0.592	55.8	-0.657	0.811	58.6	-0.873	1.015	25.9
		48	0.283	0.535	64.1	-0.655	0.83	67.0	-0.908	1.053	30.6
		68	0.341	0.566	66.7	-0.708	0.859	71.2	-0.842	0.991	33.7
	500	28	0.388	0.622	67.0	-0.615	0.768	83.7	-0.682	0.850	46.9
		48	0.356	0.589	61.0	-0.590	0.754	91.4	-0.701	0.877	54.1
		68	0.223	0.596	45.1	-0.531	0.696	92.0	-0.702	0.884	60.3
	1000	28	0.254	0.566	60.8	-0.541	0.694	95.1	-0.680	0.848	58.0
		48	0.112	0.573	38.1	-0.532	0.688	96.5	-0.658	0.823	66.8
		68	0.113	0.585	34.1	-0.426	0.576	84.1	-0.623	0.793	73.8
Under-specified Model	100	28	0.232	0.541	64.7	-0.758	0.918	61.0	-0.812	0.970	37.6
		48	0.375	0.604	69.5	-0.690	0.831	67.2	-0.754	0.903	43.1
		68	0.346	0.523	82.0	-0.584	0.753	77.5	-0.785	0.932	50.5
	500	28	0.442	0.656	76.4	-0.595	0.775	91.0	-0.679	0.862	65.2
		48	0.354	0.570	73.0	-0.528	0.742	100.0	-0.615	0.806	79.0
		68	0.247	0.530	70.0	-0.450	0.649	100.0	-0.551	0.770	84.3
	1000	28	0.509	0.707	74.5	-0.527	0.748	100.0	-0.562	0.768	88.0
		48	0.329	0.591	63.4	-0.394	0.606	96.6	-0.525	0.727	92.2
		68	0.291	0.621	59.3	-0.302	0.575	89.2	-0.535	0.741	99.4
Linear Model	100	28	0.357	0.602	63.8	-0.766	0.890	51.6			
		48	0.407	0.63	72.7	-0.624	0.799	56.0			
		68	0.409	0.624	77.1	-0.683	0.821	59.1			
	500	28	0.409	0.635	74.4	-0.565	0.768	87.1			
		48	0.383	0.640	74.1	-0.561	0.769	88.7			
		68	0.352	0.624	69.8	-0.558	0.767	81.9			
	1000	28	0.394	0.663	70.9	-0.529	0.783	98.5			
		48	0.342	0.616	61.0	-0.527	0.790	90.3			
		68	0.343	0.656	62.9	-0.660	0.896	86.0			

Note. RMSE=root mean squared error; Percent=percentage of 95% intervals, computed using the 2.5% and 97.5% thresholds of credible interval, covering the true values.

Results for Attribute Main Effect Parameter

Across all models and conditions, attribute main effect estimates were all negative, which means that they were likely underestimated in our case. For the correct model, bias and RMSE decrease when test length increases, holding sample size in control. Likewise, when sample size goes up, bias and RMSE goes down if holding test length in constant. In short, the large sample size combining the long test length seem to generate a more accurate attribute main effect estimation for the hypothetically correct model.

In terms of the over-specified model, it is interesting to note that bias and RMSE values go up when the test length increases within the small sample size, which indicates that the conditions with small sample size are not ideal for the over-specified model. When the sample size is medium to large, bias and RMSE values drop when the test length goes up, indicating that the more complex conditions are in the over-specified model, the more accurate the main effect estimation will be. Moreover, the best condition in this case would be the ones where the sample size is medium to large and the test length is medium to long.

On the other hand, the under-specified model owns the least magnitude of bias values comparing to other models. This leads us to suspect that the under-specified model is the best model for attribute main effect estimation. Again, the bias and RMSE values react to sample size and test length as we expected. Holding sample size in constant, the bias and RMSE decrease as the test length goes up. Likewise, the bias and RMSE decrease as the sample size increases if keeping test length in control. In short, based on what we see in terms of bias and RMSE values, the under-specified model is the most accurate model in estimating attribute main effect, with the a more complex condition – large sample size combining with long test length ideally.

Lastly, the linear model again yielded the greatest magnitude of bias values in attribute main effect estimation, which confirms what we anticipated that the linear model is the wrong model in a multidimensional framework. In terms of the sample size and test length effect, the linear model follows the same pattern as shown in other models – as the sample size goes up, the bias and RMSE go down if holding test length in constant; and as the test length goes up, the bias and RMSE values decrease when keeping sample size in control. For this model, the best condition is when the sample size is large and the test length is medium.

In short, with the separate examination in bias and RMSE, the 95% credible intervals across models and conditions covered up to 100% of the true values. This means that the credible interval endorsed the conclusion that main effect parameters were reliably estimated in LCDM and Bayesian Networks using the MCMC algorithm.

Results for Attribute Interaction Parameter

Attribute interaction parameters occurred in the hypothetically correct model, the over-specified model, and the under-specified model, where attribute *RU* depends on at least two other attributes. With a quick glance, the attribute interaction parameter estimates were all negative as well, meaning attribute interaction parameters were likely underestimated across models and conditions. The attribute interaction terms were estimated along with attribute intercept and main effect. Interaction was not estimated in the linear model, because there is no interaction between the attributes when they located in a linear structure. Across models and conditions, bias and RMSE values are large, which can be a result of the limited number of attributes (four attributes in our case) in the model structures.

In terms of sample size and test length effect, the bias and RMSE values decrease as the sample size increases across all three models, holding test length in constant. Likewise, the bias

and RMSE goes down as the test length goes up when keeping sample size in control. The best model in comparing the bias and RMSE values among the three models in estimating attribute interaction parameters seems to be the under-specified model, who has the least magnitude of bias and RMSE in general. And the best condition in this case is when the sample is large and the test length is medium. Besides, the interaction terms been estimated in the three models — correct model, over-specified model, and under-specified model—all provided reliable results in terms of parameters estimates. Moreover, the credible interval covers up to 99.4% of the true values, suggesting that these attribute interaction parameters have been well estimated, and therefore the results are reliable.

In order to quantify the effect of sample size, test length and model specification, an ANOVA was conducted, and the results were presented in Table 4.6. As shown in Table 4.6, the slightly bigger effect on attribute parameter recovery is model specification, especially on attribute main effects. Test length almost does not have any effect on attribute parameter recovery, and sample size just has small effect on attribute main effects and attribute interactions. This result indicates that how a map is structured, or how a set of skills was connected in an assessment, does impact students' mastery status of the skills.

Table 4.6.

Eta-Squared for Attribute Parameter Recovery by Condition

		Sample Size	Test Length	Model Specification
Bias	Intercept	0	0	0.01
	Main Effect	0.06	0.01	0.01
	Interaction	0.08	0	0.02
RMSE	Intercept	0	0	0
	Main Effect	0.05	0.01	0.10
	Interaction	0.07	0	0.02

Note. Rule of thumb for η^2: 0.02=small, 0.13=medium, 0.26=large (Cohen, 1998; Miles & Shevlin, 2001)

After examining the attribute parameter accuracy and finding consistently reliable results in the parameter estimation, the impact of the models and conditions to student mastery classification status can then be explored as a further step. The next section presents the estimates of respondents' mastery status towards the four attributes and the test as a whole by examining the accuracy of respondents' skill estimation.

Student Mastery Classification Accuracy

This section discusses the student mastery classification calculated based upon true values and estimated data. Table 4.7 below presents a concise summary of student mastery classification in percent. As shown in the table, the over-specified model has the highest classification rate overall with 68.43%, and the linear model has the lowest rate, 66.29%. As per attribute, attribute *APP* is not dependent on any other attributes, and the classification rate is similar across all four models with the highest percentage (67.14%) in the over-specified model and the lowest (63.90%) in the linear model; the correct model and the under-specified model have a very similar mastery classification rate, 66.24% and 66.95% respectively.

For attribute *PI*, the over- specified model again has the highest classification rate, followed by the under-specified model and the correct model. It is worth noting that attribute *PI* in the over-specified model has a parent attribute, *APP*, and with this dependency upon attribute *APP*, the mastery classification of attribute *PI* raised very slightly compared to those without a parent attribute. This finding confirms that there is a hierarchy among the attributes, and by mastering one attribute increases the probability of mastering the child attribute, *PI* in our case. On the other hand, the mastery status in percentage for attribute *MC* is very similar across all four models, even in the linear model, where attribute *MC* is dependent upon attribute *PI*. This leads us to believe that the linear model again is not the correct model, because the mastery of

MC does not show a significantly higher rate compared to the rest of the models where attribute *MC* does not have any parent attribute. In other words, students are supposed to have a better chance mastering attribute *MC* when *MC* is dependent upon another attribute, *PI* in our case, but in the linear model, the mastery of *PI* does not improve students' mastery of attribute *MC*.

Finally, when we come to attribute *RU*, it is noticeable that the student mastery classification rate is higher when there are more parent attributes linked to it. As suspected, the linear model has the lowest mastery rate for attribute *RU* because there is only one attribute predicting the probability of *RU* mastery. The correct model and over-specified model both own similar rate in this sense, both of which have three attributes predicting attribute *RU*. Although, the rate is slightly higher in the over-specified model. It is interesting to note that in the under-specified model, attribute *RU* depends on two parent attributes, and yet the mastery rate was even 0.2% higher, though negligible, from those dependent on three. To summarize, by increasing the dependency to an attribute, the mastery classification rate to that attribute increases slightly as well, though not significantly, indicating the attribute structure exists.

Table 4.7

Mean student classification by estimation in percentage

Model	APP	PI	MC	RU	Overall
Hypothetically Correct Model	66.24	67.55	65.41	72.13	67.84
Over-specified Model	67.14	68.38	65.92	72.27	68.43
Under-specified Model	66.95	68.11	65.02	72.47	68.14
Linear Model	63.90	66.28	65.07	69.89	66.29

Note. Percentages are calculated based on the estimated student mastery status

Moreover, Table 4.8 describes the student mastery classification of estimated data for each attribute and the hypothetical overall test across conditions. When looking at the difference between the estimated mastery classification rates in the hypothetically correct model versus in

the rest of the models, the under-specified model is the closest in student mastery classification for the overall mastery of the test, followed by the over-specified model. The linear model, on the other hand, tends to underestimate student mastery status, and thus has the largest difference in student mastery classification from the correct model. Another interesting finding is that attribute *RU* was estimated more accurately by the under-specified model when comparing with the results from the over-specified model. The linear model owns the lowest mastery classification rate, for each of the four attributes, and thereby is an indication of an incorrect model in this case. Furthermore, based on the least magnitude of the differences between estimated mastery rate in the hypothetically correct model versus in the other models, attributes with fewer than one parent attribute can be estimated more accurately when in long test length, whereas attributes that have more than one parent attribute seem to be better estimated with short test length. In terms of the sample size influence, students can be more accurately categorized as mastery/non-mastery as the sample size of the whole assessment increases, holding test length in constant. The same principle applies to the test length effect if holding sample size in control.

Student mastery classification was calculated per attribute as well as the overall test. The mastery rate was also compared with the rate generated by the correct model. The results indicate that as the parent attribute exists or increases, student mastery toward the child attribute increases as well. This is congruent to the hypothesis made in the beginning of this study that there is a hierarchy among the attributes, and by exploring the possible structures of the attributes we might be able to pinpoint the best structure that makes the most accurate estimates and student mastery classifications.

Table 4.8

Detailed student mastery classification in percentage

Condition			APP	PI	MC	RU	Overall
Model	Sample Size	Test Length					
Correct Model	100	28	69.25	75.41	71.11	88.69	76.12
		48	72.40	74.28	67.71	88.23	75.66
		68	70.47	72.89	65.03	88.35	74.19
	500	28	70.33	70.65	69.57	69.29	69.96
		48	66.31	65.78	65.23	65.94	65.82
		68	61.28	61.40	61.02	60.86	61.14
	1000	28	66.81	68.07	67.94	68.00	67.71
		48	62.00	61.86	63.36	62.03	62.31
		68	57.34	57.60	57.76	57.75	57.61
Over-specified Model	100	28	72.42	78.96	72.11	88.84	78.08
		48	74.13	78.03	69.15	88.52	77.46
		68	72.72	74.59	65.10	87.71	75.03
	500	28	70.97	70.81	71.62	72.01	71.35
		48	64.47	64.09	64.70	64.46	64.43
		68	60.33	60.13	60.50	60.70	60.42
	1000	28	70.22	70.28	70.47	69.57	70.14
		48	61.46	60.94	61.37	60.68	61.11
		68	57.57	57.62	58.23	57.96	57.85
Under-specified Model	100	28	70.99	76.09	71.42	89.58	77.02
		48	71.46	75.39	65.75	86.91	74.88
		68	73.23	74.71	63.51	88.75	75.05
	500	28	68.92	69.22	68.42	68.64	68.80
		48	64.69	65.07	65.03	65.20	65.00
		68	63.81	63.85	63.82	63.87	63.84
	1000	28	69.30	69.35	68.54	68.54	68.93
		48	61.61	61.09	60.29	62.21	61.30
		68	58.57	58.24	58.41	58.49	58.43
Linear Model	100	28	66.01	78.89	78.20	87.18	77.57
		48	70.70	77.21	71.94	86.35	76.55
		68	69.19	74.23	68.05	86.08	74.39
	500	28	65.35	63.82	65.67	65.92	65.19
		48	64.66	64.49	64.09	64.25	64.37
		68	60.94	60.27	59.90	60.53	60.41
	1000	28	64.95	63.78	64.21	64.49	64.36
		48	57.72	57.68	58.19	58.63	58.06
		68	55.58	56.16	55.42	55.61	55.69

CHAPTER FIVE

Real Data Analysis Results

The data used in this study is from the Diagnosing Teachers' Multiplicative Reasoning (DTMR) test conducted by Bradshaw, Izsak, Templin and Jacobson (2014), which was built within a diagnostic classification framework. The prior research focused on the diagnostic quality of the test at the item level, mastery classifications for teachers, and attribute relationships. It successfully demonstrated that the Diagnostic Classification Models (DCMs) can detect distinct patterns of attribute mastery. However, the hierarchies among the attributes were not examined. Since attribute hierarchies were not explored in the original study, this study could be viewed as an expansion of the prior study by Bradshaw, et al. (2014). This research analyzed data from the 974 teachers who took the DTMR test and responded to 27 items measuring four attributes.

In light of the hierarchical features found in the data based on the previous research, this study further explored possible hierarchical structures among the attributes by using four models: the hypothetically correct model (Figure 3.2), the over-specified model (Figure 3.3), the under-specified model (Figure 3.4), and the linear model (Figure 3.5). For each of the four models, a proficiency model on the attribute level and person mastery level, and an evidence model on the item level, were included in the estimation. The proficiency model, created according to the four possible attribute structures, was shown in Figure 3.2 to Figure 3.5. The number of items within each attribute, used in the evidence model, was presented in Figure 3.1. Probabilities in both the proficiency model and the evidence model for each corresponding attribute structure were calculated using Loglinear Cognitive Diagnosis Model (LCDM) integrated with Bayesian Networks. The MCMC estimation was employed for both attribute level and item level

parameter estimation. In the MCMC estimation, two chains were used, each of which has a length of 5,000, and the first 1,000 iterations were discarded as burn-ins. The model convergence and parameter estimation results for each attribute structure are presented in the following sections.

Model Fit and Model Convergence

The goodness-of-fit of the models were examined using the deviance. The deviance presented in Table 5.1 was directly reported by JAGS. The smaller the deviance is, the better the model fits the data. Based on this principle, the misspecified model turned out to be the best-fitting model for the DTMR test data, followed by the over-specified model, the under-specified model, and the correct model. Again, the reason is that the misspecified model is the least complex model, and thus was penalized the least, which yielded the smallest DIC among the four models. In addition to the DIC values, attribute and item parameter recovery should be also be examined while considering the best fitting model to the data.

R-hat was commonly used as a measure of model convergence for the MCMC algorithm in the Bayesian Network. An R-hat value close to 1 indicates that the model converges well as a whole, and the percentage of R-hat values less than 1.2 was counted as a good evidence of convergence (recommended by Gelman, 1996) or 1.1 (default value of JAGS). Table 5.1 presents the percentage of R-hat values greater than 1.2 for each parameter within each model, along with the corresponding maximum R-hat values. Convergence was examined, using R-hat, in three levels: (1) Overall model fit - the deviance; (2) Proficiency model fit - attribute parameters and person proficiency variable; (3) Evidence model fit - item parameters.

First, convergence at the model level was examined. As shown in the left section of Table 5.1, the R-hat values of the deviance for the all models were very close to 1, indicating that the estimation at the model level converged well in terms of R-hat values of the deviance.

Next, the convergence for attribute parameters—intercept, main effect, and interaction—were all good across models, by looking at the percentage of R-hat values less than 1.2. Finally, the convergence at the evidence level was considered, and none of the item parameters showed a value greater than 1.2. Therefore, all four models showed good convergence, with close to 1 R-hat values for item intercept, item main effect, and item interaction. More detailed findings on the attribute level, item level, and mastery classification rate for each attribute structure are elaborated in the following sections.

Table 5.1

The percentage of R-hat values greater than 1.2 and the maximum R-hat for each parameter set

Model	Deviance		Attribute						Item					
			Intercept		Main Effect		Interaction		Intercept		Main Effect		Interaction	
	Value	R-hat	%	Max	%	Max	%	Max	%	Max	%	Max	%	Max
Hypothetically Correct Model	27462.36	1.00	0	1.01	0	1.00	0	1.00	0	1.02	0	1.02	0	1.01
Over-specified Model	27420.05	1.00	0	1.03	0	1.01	0	1.00	0	1.01	0	1.02	0	1.01
Under-specified Model	27453.24	1.00	0	1.01	0	1.00	0	1.00	0	1.01	0	1.03	0	1.03
Linear Model	27383.02	1.01	0	1.01	0	1.01	-	-	0	1.02	0	1.02	0	1.01

Results for Item Parameters

The item parameter estimates for the four models: the correct model, the over-specified model, the under-specified model, and the linear model, are provided in Table 5.2 through Table 5.5. Item parameter estimates consist of estimates for item intercept, item main effect and item interaction, for individual item. Across the four models, the average item intercept ranged from −.71 to −.72, meaning just 7.1% to 7.2% of respondents who had not mastered any of the four attributes (*APP*, *PI*, *MC*, and *RU*) answered the items correctly. The average main effect parameters ranged from .74 to 1.22, and the average interaction parameter estimates fell between 0.53 and 1.05. In general, items with lower intercepts and higher main effects and interaction terms are more discriminating between masters and non-masters of the attributes. In this sense, there are some items that are consistently showing high main effects and interactions across models, especially for item 8b, 8c, 9, 10b, and item 15c, and thus are more discriminating than other items.

Table 5.2

Item parameter estimates for the hypothetically correct model

		RU(α_4)	APP(α_1)	PI(α_2)	MC(α_3)	PI/MC	RU/MC	RU/PI
i	$\lambda_{i,0}$	$\lambda_{i,1(4)}$	$\lambda_{i,1(1)}$	$\lambda_{i,1(2)}$	$\lambda_{i,1(3)}$	$\lambda_{i,2(2,3)}$	$\lambda_{i,3(3,4)}$	$\lambda_{i,2(2,4)}$
1	0.60				0.19			
2	−0.98	0.96						
3	0.86		0.92					
4	−1.40				0.94			
5	−1.08				0.34			
6	−1.41		0.76					
7	−2.31				0.50			
8a	−0.57	0.36						
8b	−0.05	2.14						
8c	0.12	1.66						
8d	0.36	0.67						
9	−0.42				1.27			
10a	−1.02			0.14				
10b	0.27			0.27	1.49		0.97	
10c	−1.27			0.11	1.30		1.13	
11	−1.37				1.70			
12	−0.76				0.69			
13	−1.23		0.58	0.12		0.75		
14	−0.56		0.22		0.28			0.26
15a	−1.83		0.33	0.07		0.38		
15b	−0.98		1.92					
15c	0.05		1.76					
16	0.23				1.55			
17	−0.80		0.27		0.32			0.54
18	−1.54		0.39		0.58			0.68
21	−1.28				0.96			
22	−0.92		0.79		0.72			0.76
Average	−0.71	1.16	0.79	0.14	0.86	0.57	1.05	0.56

Table 5.3

Item parameter estimates for the over-specified model

		RU(α_4)	APP(α_1)	PI(α_2)	MC(α_3)	PI/MC	RU/MC	RU/PI
i	$\lambda_{i,0}$	$\lambda_{i,1(4)}$	$\lambda_{i,1(1)}$	$\lambda_{i,1(2)}$	$\lambda_{i,1(3)}$	$\lambda_{i,2(2,3)}$	$\lambda_{i,3(3,4)}$	$\lambda_{i,2(2,4)}$
1	0.59				0.19			
2	−1.09	1.29						
3	0.85		0.92					
4	−1.41				0.96			
5	−1.09				0.35			
6	−1.45		0.81					
7	−2.32				0.52			
8a	−0.62	0.51						
8b	0.16	2.09						
8c	0.34	1.45						
8d	0.37	0.78						
9	−0.44				1.28			
10a	−1.01			0.14				
10b	0.27			0.26	1.46		0.98	
10c	−1.29			0.12	1.28		1.14	
11	−1.39				1.68			
12	−0.77				0.70			
13	−1.26		0.57	0.12		0.78		
14	−0.57		0.23		0.27			0.26
15a	−1.85		0.35	0.07		0.40		
15b	−1.03		1.96					
15c	0.02		1.76					
16	0.21				1.56			
17	−0.81		0.26		0.31			0.51
18	−1.55		0.40		0.57			0.64
21	−1.29				0.96			
22	−0.94		0.82		0.63			0.77
Average	−0.72	1.22	0.81	0.14	0.85	0.59	1.06	0.54

Table 5.4

Item parameter estimates for the under-specified model

		RU(α_4)	APP(α_1)	PI(α_2)	MC(α_3)	PI/MC	RU/MC	RU/PI
i	$\lambda_{i,0}$	$\lambda_{i,1(4)}$	$\lambda_{i,1(1)}$	$\lambda_{i,1(2)}$	$\lambda_{i,1(3)}$	$\lambda_{i,2(2,3)}$	$\lambda_{i,3(3,4)}$	$\lambda_{i,2(2,4)}$
1	0.59				0.21			
2	−0.96	0.95						
3	0.87		0.92					
4	−1.40				0.93			
5	−1.08				0.34			
6	−1.41		0.76					
7	−2.31				0.50			
8a	−0.58	0.38						
8b	−0.03	2.13						
8c	0.12	1.67						
8d	0.35	0.71						
9	−0.43				1.27			
10a	−1.01			0.15				
10b	0.29			0.27	1.61		0.90	
10c	−1.28			0.13	1.43		1.07	
11	−1.40				1.72			
12	−0.77				0.69			
13	−1.22		0.64	0.14		0.71		
14	−0.56		0.22		0.28			0.27
15a	−1.82		0.36	0.08		0.37		
15b	−0.97		1.91					
15c	0.05		1.78					
16	0.23				1.53			
17	−0.80		0.27		0.32			0.55
18	−1.56		0.39		0.59			0.70
21	−1.28				0.94			
22	−0.94		0.82		0.71			0.77
Average	−0.71	1.17	0.81	0.15	0.87	0.54	0.98	0.57

Table 5.5

Item parameter estimates for the linear model

		RU(α_4)	APP(α_1)	PI(α_2)	MC(α_3)	PI/MC	RU/MC	RU/PI
i	$\lambda_{i,0}$	$\lambda_{i,1(4)}$	$\lambda_{i,1(1)}$	$\lambda_{i,1(2)}$	$\lambda_{i,1(3)}$	$\lambda_{i,2(2,3)}$	$\lambda_{i,3(3,4)}$	$\lambda_{i,2(2,4)}$
1	0.57				0.25			
2	−1.01	1.13						
3	0.83		0.94					
4	−1.36				0.85			
5	−1.06				0.30			
6	−1.45		0.80					
7	−2.28				0.44			
8a	−0.59	0.46						
8b	0.14	2.09						
8c	0.30	1.52						
8d	0.34	0.82						
9	−0.42				1.24			
10a	−1.07			0.32				
10b	0.23			0.96	1.26		0.83	
10c	−1.44			0.80	0.98		1.06	
11	−1.42				1.75			
12	−0.75				0.66			
13	−1.28		0.33	0.51		0.66		
14	−0.58		0.26		0.27			0.27
15a	−1.86		0.25	0.20		0.40		
15b	−1.03		1.91					
15c	0.00		1.77					
16	0.27				1.38			
17	−0.81		0.22		0.33			0.66
18	−1.54		0.44		0.53			0.72
21	−1.25				0.88			
22	−0.98		0.92		0.64			0.85
Average	−0.72	1.20	0.78	0.56	0.79	0.53	0.94	0.62

As far as mastery status for each individual item is concerned, an item characteristic bar chart plots the response probabilities on the vertical axis as a function of attribute mastery on the horizontal axis. Figure 5.1 through Figure 5.7 present characteristics of simple structure items (measuring one attribute only) and complex structure items (measuring two attributes). Figure 5.1 through Figure 5.4 present the ICBCs for items measuring only one attribute. Take Figure 5.1 for an example, in the hypothetically correct model, the probability of answering *Item 2* correctly was .5 for masters of *APP* and .27 for non-masters. Figure 5.5 through Figure 5.7 depict the ICBCs for items measuring two attributes. When an item measures two attributes, the interaction effect between the two attributes were also accounted for and examined in the item estimated probability. For instance, *Item 22* in the hypothetically correct model, measures both attribute *PI* and *RU*, the probability of a correct response to *Item 22* increases from .28 to .45 to .47 to .79 when comparing students who mastered neither attributes, only *RU*, only *PI*, and both attributes, respectively.

Furthermore, as discussed in the previous chapter, model specification or attribute structure does not have much effect on item level estimates. The similar findings were also seen in the real data analysis results, as shown in the item characteristic bar charts (ICBCs) in Figure 5.1 through Figure 5.7. For example, for items measure attribute *APP* only (Figure 5.1), the estimated probabilities of answering *Item 2* correctly as masters of *APP* were .50 for the hypothetically correct model, .55 for the over-specified model, .50 for the under-specified model, and .53 for the linear model, comparing to the non-masters with probabilities .27, .25, .28, and .27, respectively. Even for the complex structure items – items that measure two attributes, the results were similar. Take Figure 5.7 for an example, *Item 14* measure *PI* and *RU*, and the probabilities of answering *Item 14* correctly were close to .55, .42, .43, .36 for students who

mastered both attributes, only *PI*, only *RU*, and neither attributes, respectively, across four models.

Therefore, the variation on the proficiency level − attribute structure change − does not influence the mastery probability of items. One of the reasons is that the variation in the attribute structure was small, with only the addition or removal of one random pathways. The number of attributes are also very limited, and thereby the influence of the attribute structure to the correctness of item responses could not be fully explored. Moreover, the number of items within each attribute might not be large enough to capture all variations in the mastery of items as the attribute structure changes. For instance, there was just one item measuring only attribute *MC*. This matter was explored further in the simulation study.

To sum up, students' probability of answering an item correctly or not does not rely on how the attributes were structured in a test, rather, how well the item connects to the targeted attribute. This is critical to item development in educational assessment. When items are well written and specified to measure certain attributes, even when the underlying connection of skills is not informative or even not correct, students' performance on the items will not be influenced in a negative way. More specifically, the probability of a student answering an item correctly will not improve as the dependencies among the skills increase or decrease. Although it does not mean that a wrong map structure should be used in a test development, because the specification of attributes and their connections does have impact on students' mastery classification on the test. As much as we would like more students to be classified as masters of skills in an assessment, the accuracy of that classification is critical when decision making is involved.

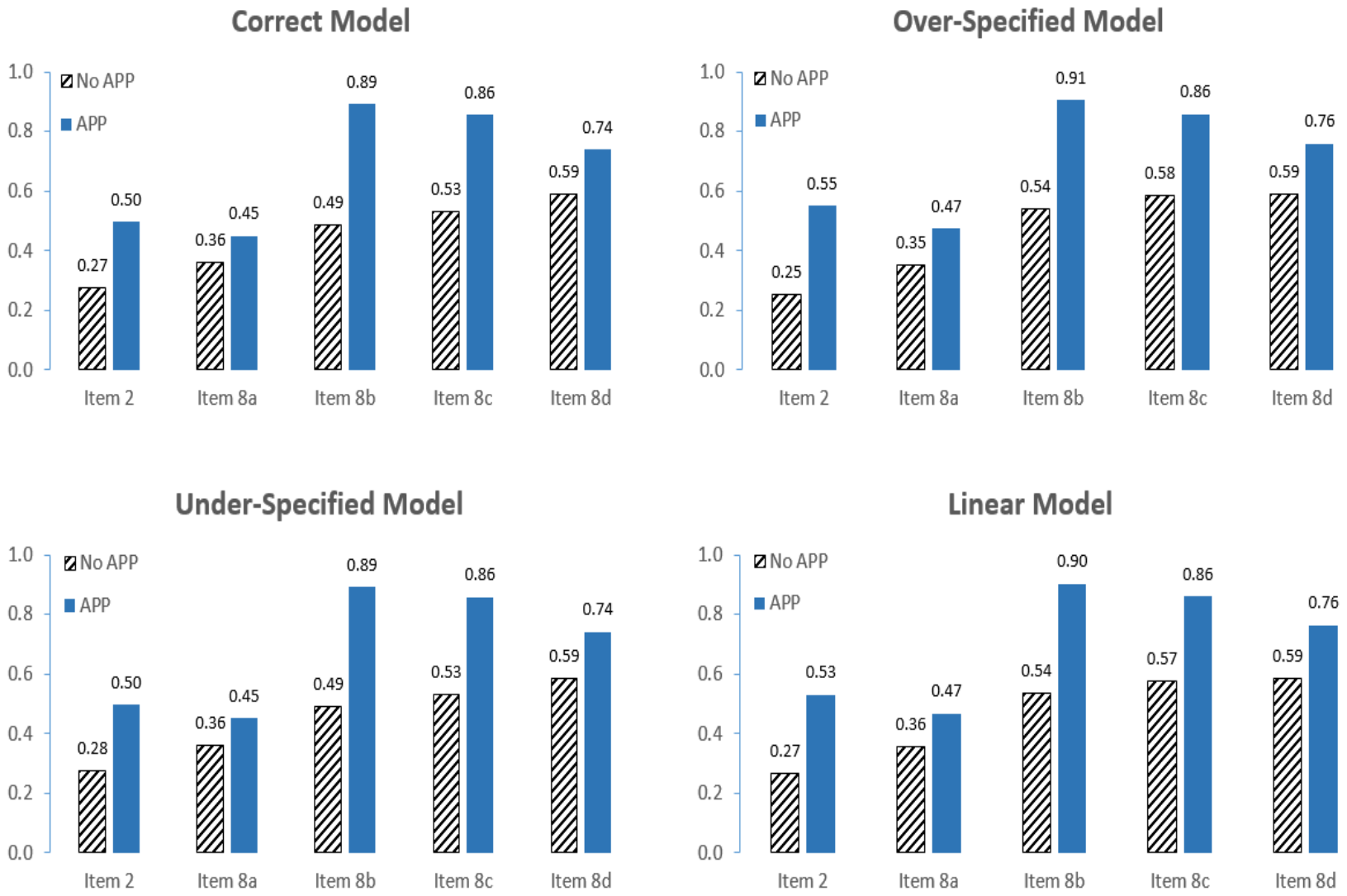

Figure 5.1 . Item characteristics bar charts for items measuring Appropriateness (APP). Five items on the DTMR test measured this attribute. The figure displays the probability of a correct response (vertical axis) by the discrete mastery state (horizontal axis).

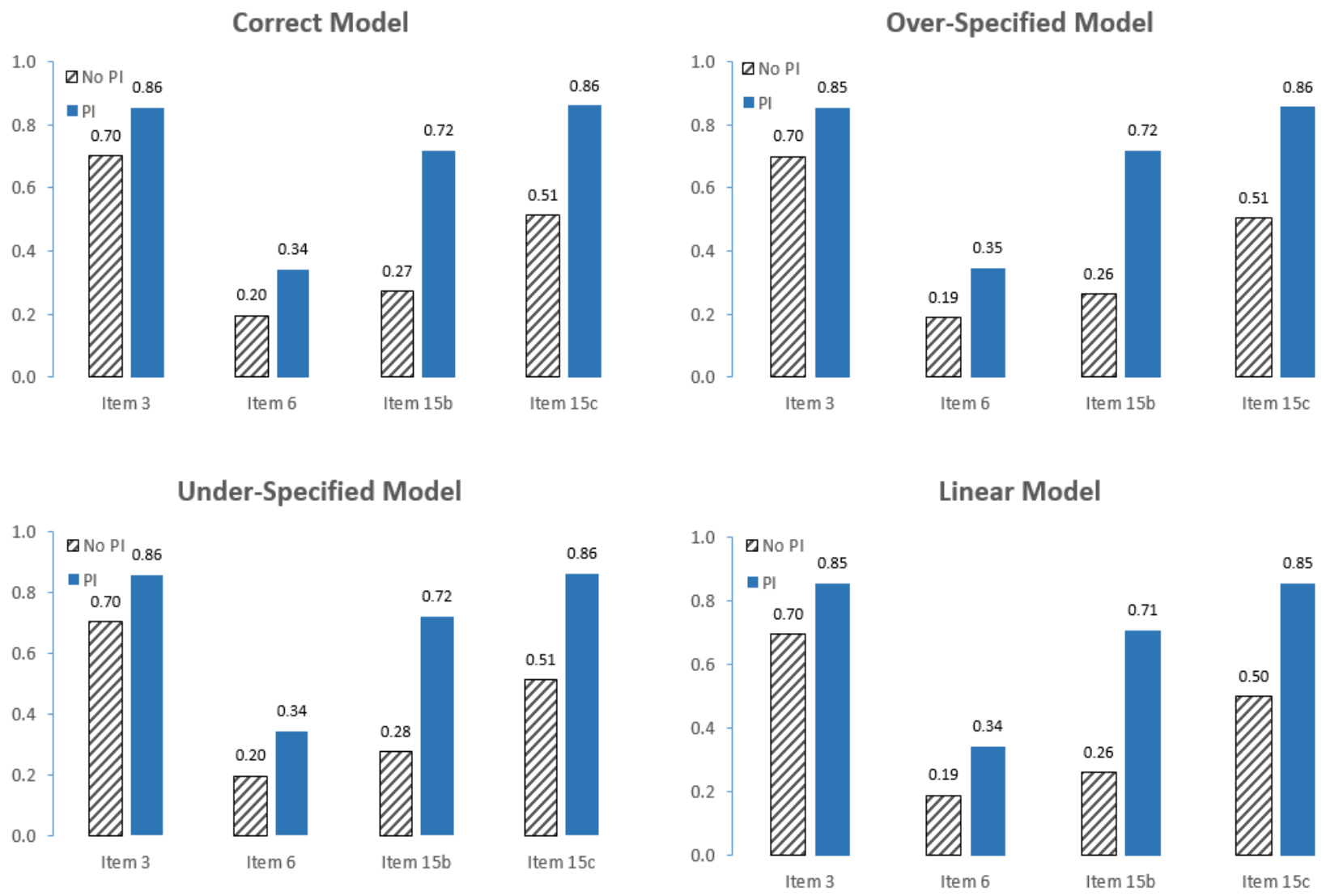

Figure 5.2. Item characteristics bar charts for items measuring Partitioning & Iterating (PI). Four items on the DTMR test measured this attribute. The figure displays the probability of a correct response (vertical axis) by the discrete mastery state (horizontal axis).

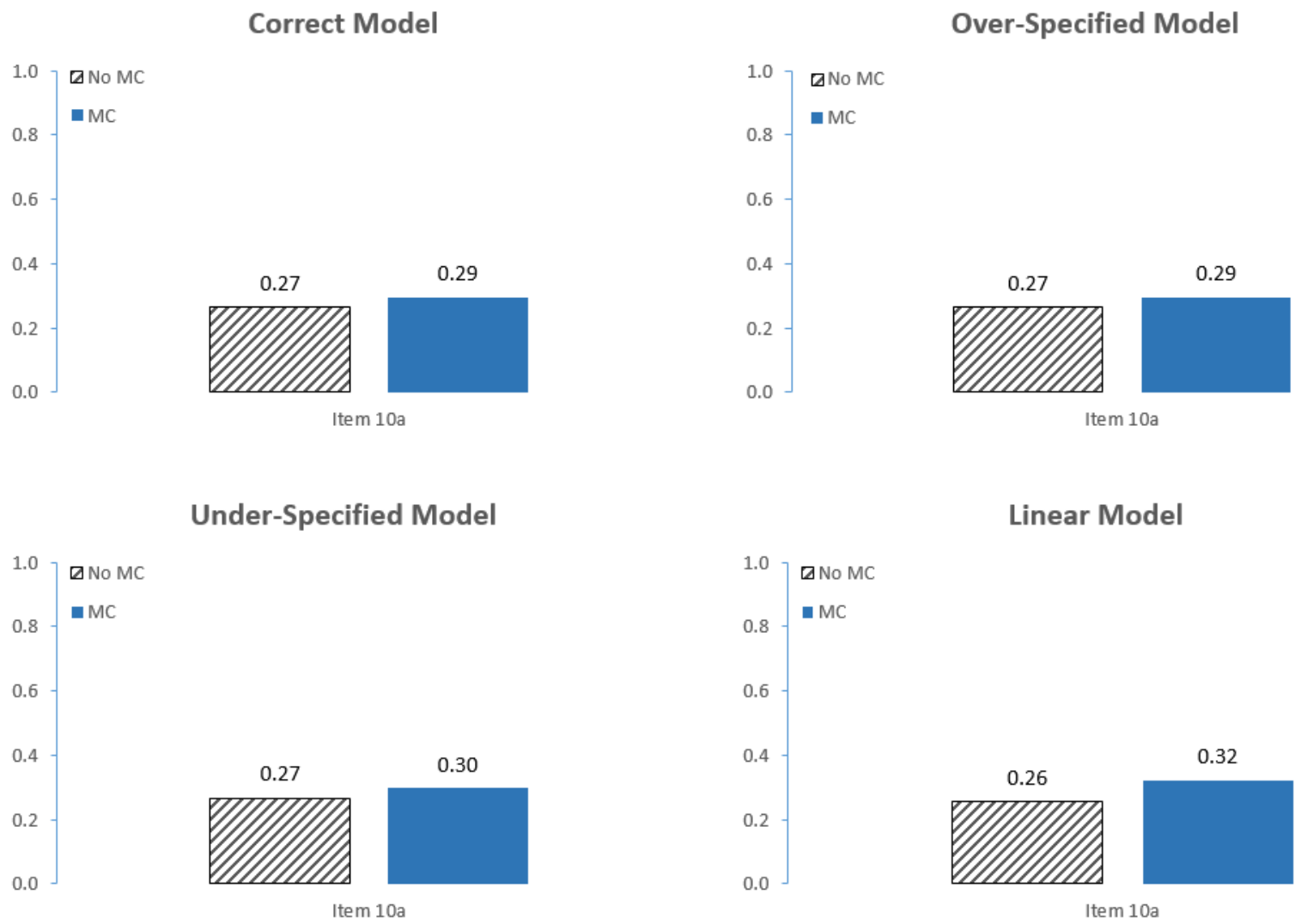

Figure 5.3. Item characteristics bar charts for items measuring Multiplicative Comparison (MC). Only one item on the DTMR test measured this attribute. The figure displays the probability of a correct response (vertical axis) by the discrete mastery state (horizontal axis).

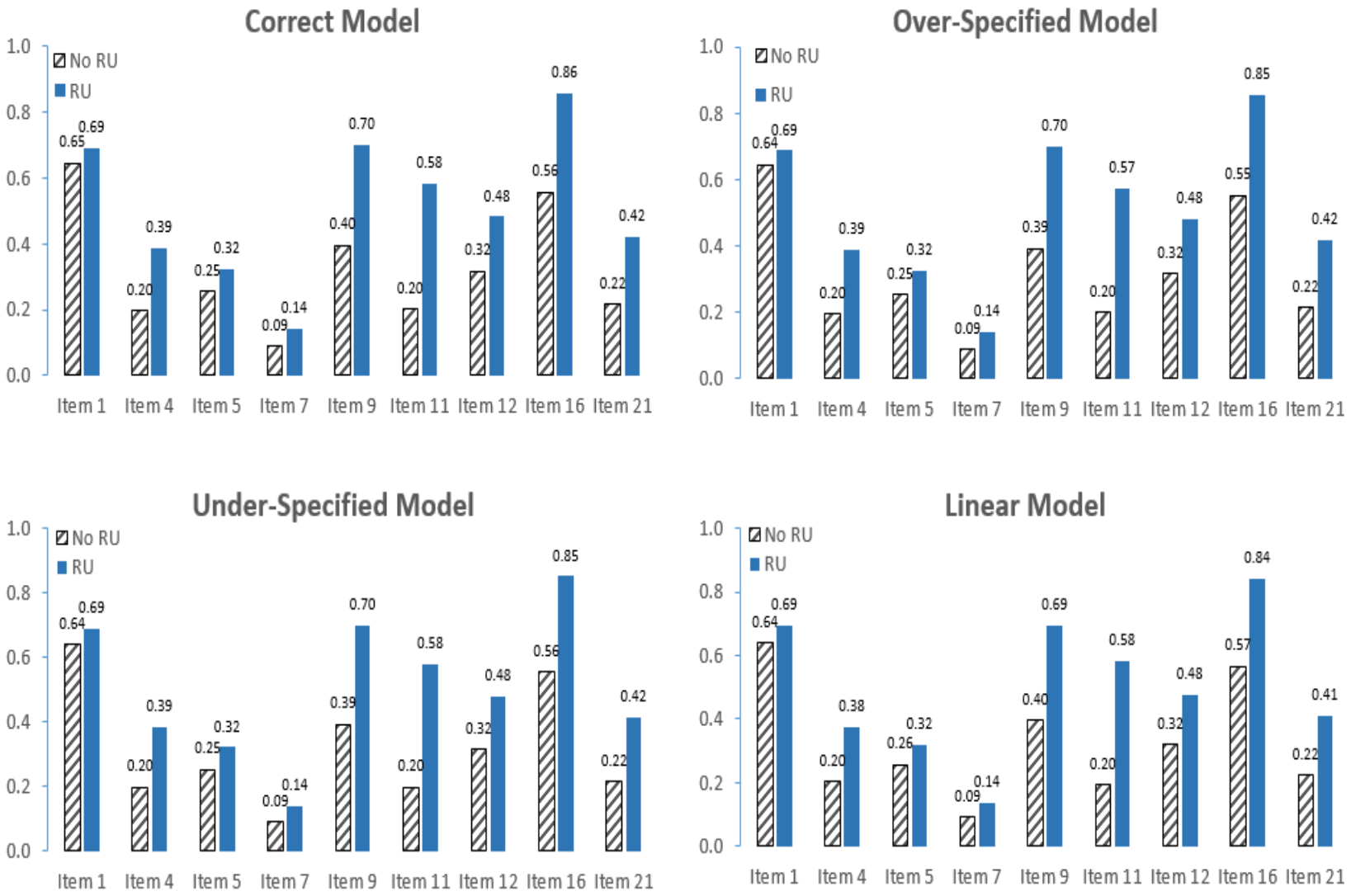

Figure 5.4. Item characteristics bar charts for items measuring Referent Unit (RU). Nine items on the DTMR test measured this attribute only. The figure displays the probability of a correct response (vertical axis) by the discrete mastery state (horizontal axis).

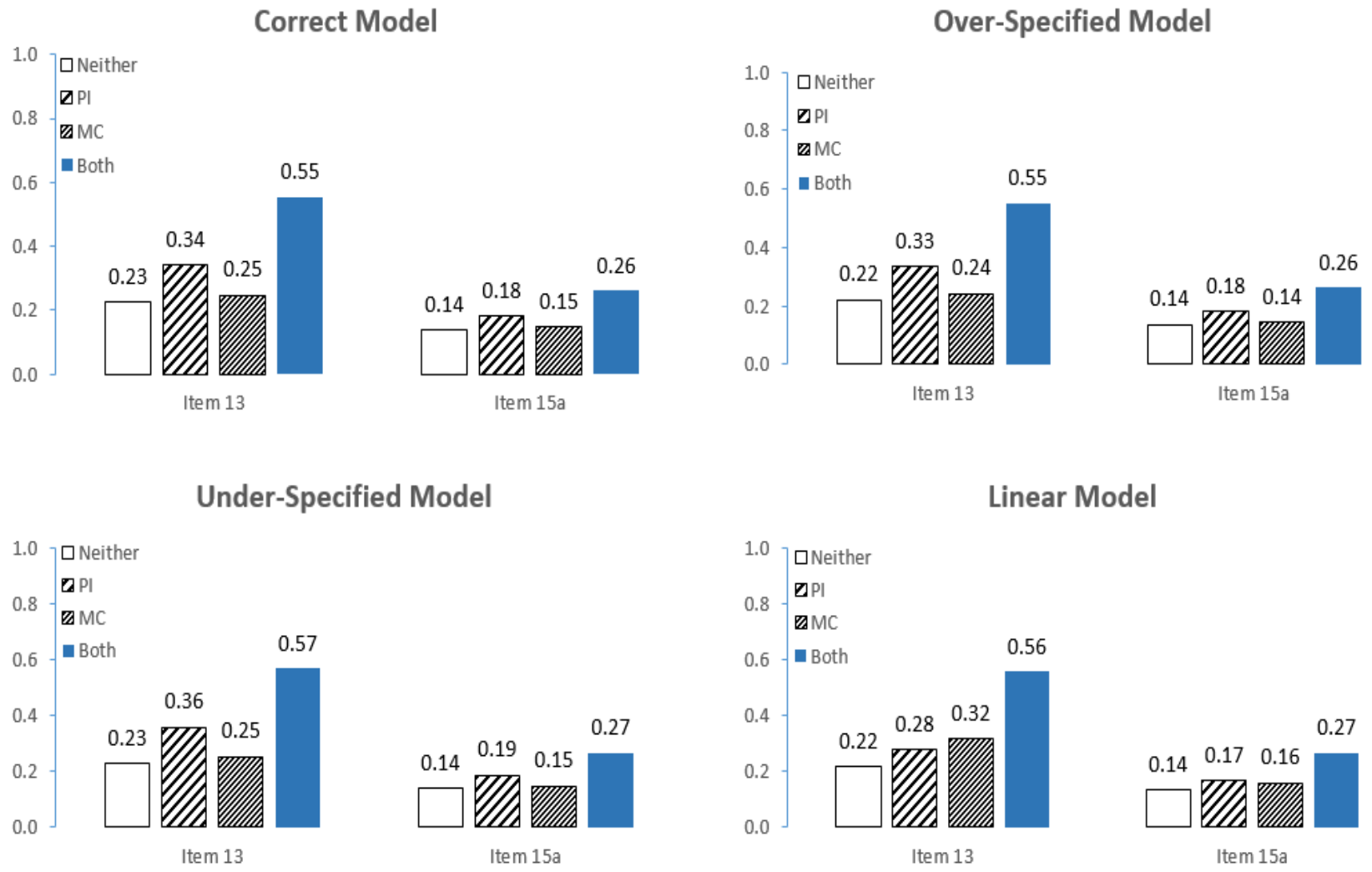

Figure 5.5. Item characteristics bar charts for items measuring Partitioning & Iterating (PI) and Multiplicative Comparison (MC). Two items on the DTMR test measured these two attributes. The figure displays the probability of a correct response (vertical axis) by the discrete mastery state (horizontal axis).

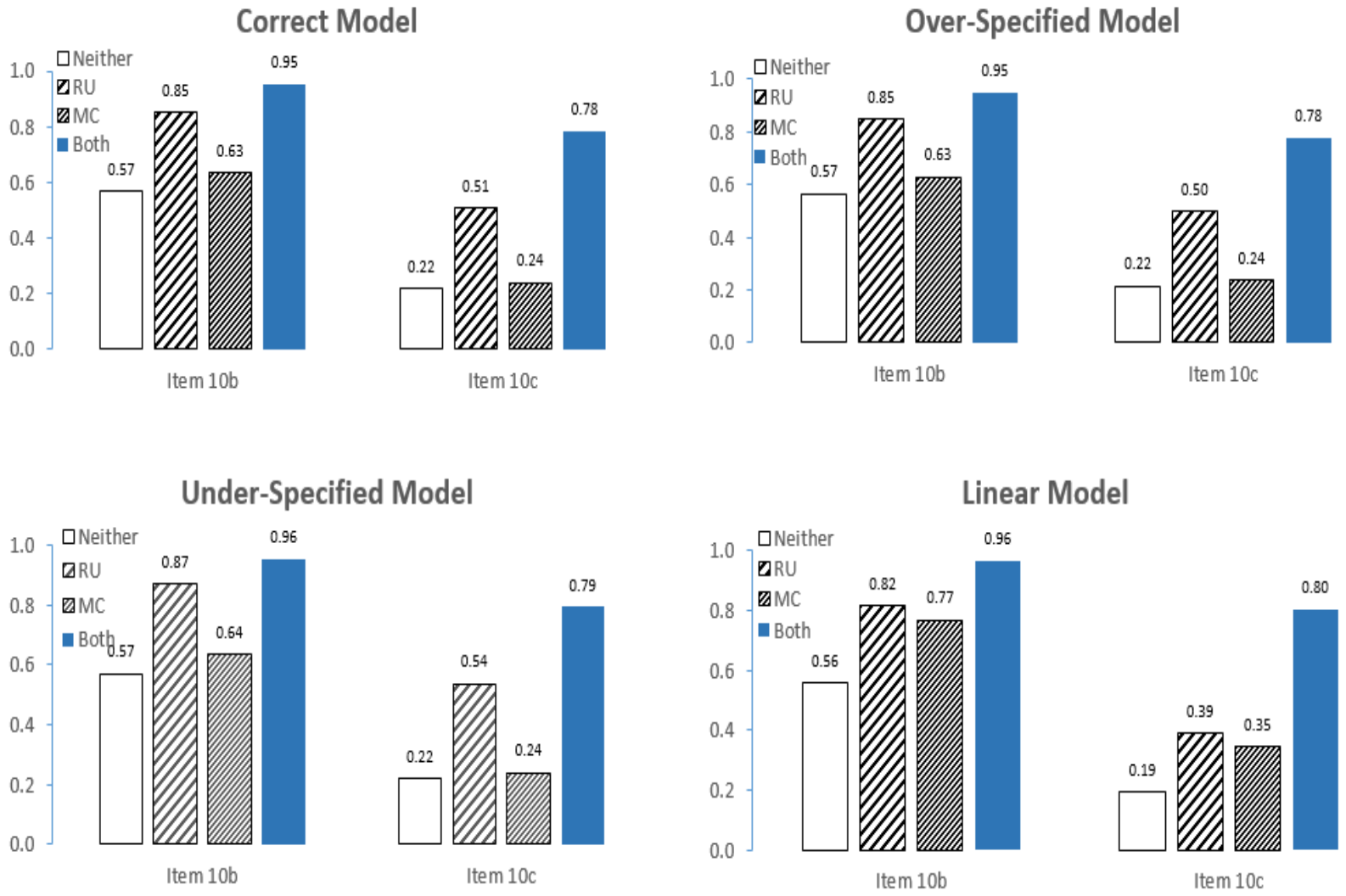

Figure 5.6. Item characteristics bar charts for items measuring Referent Unit (RU) and Multiplicative Comparison (MC). Two items on the DTMR test measured these two attributes. The figure displays the probability of a correct response (vertical axis) by the discrete mastery state (horizontal axis).

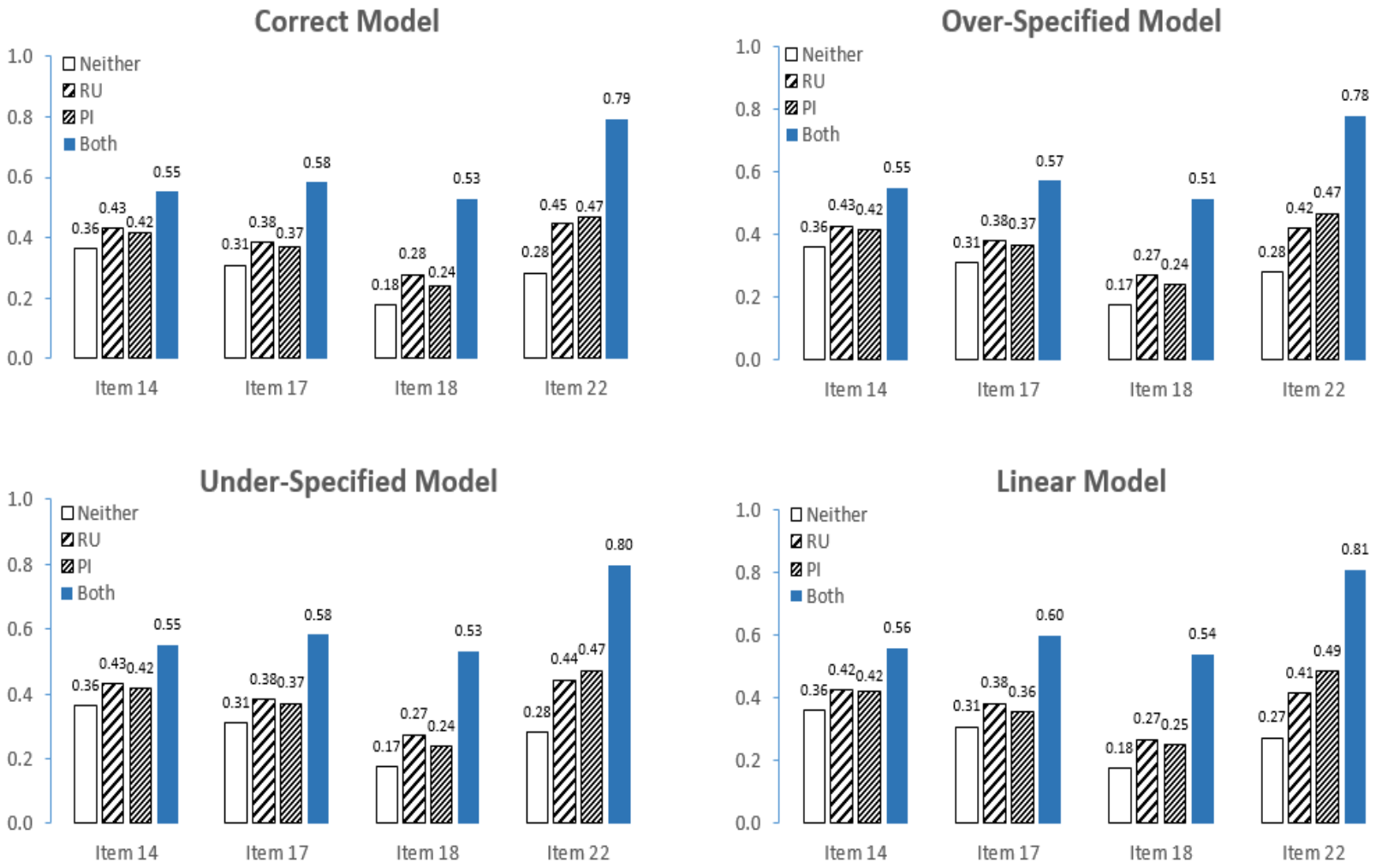

Figure 5.7. Item characteristics bar charts for items measuring Referent Unit (RU) and Partitioning & Iterating (PI). Four items on the DTMR test measured these two attributes. The figure displays the probability of a correct response (vertical axis) by the discrete mastery state (horizontal axis).

Results for Attribute Parameters

The attribute parameter estimates and the probability of each individual attribute mastery are provided in the tables below. Table 5.6 and Table 5.7 present the attribute level results from the correct model. The results showed that attribute *RU* is heavily dependent upon two attributes: *APP*, *PI*, in that the probability of *RU* mastery increases sharply as a student masters either *APP*, *PI*, or both. Surprisingly, the probability of *RU* mastery does not increase as a respondent masters attribute *MC*, suggesting that *RU* is possibly not depending on *MC* solely or the combination of *MC* and *APP*. As anticipated, when all three remaining attributes were mastered, the probability of *RU* mastery approached very close to 1.00.

Table 5.6

Mastery status of attributes for the correct model

	Logit	Probability of Mastery	Probability of Non-Mastery
APP	0.553	0.635	0.365
PI	−0.139	0.465	0.535
MC	1.021	0.735	0.265

Table 5.7

Probability of RU mastery for the correct model

APP	PI	MC	Probability of RU Mastery
0	0	0	0.214
0	0	1	0.129
0	1	0	0.417
1	0	0	0.284
0	1	1	0.459
1	0	1	0.230
1	1	0	0.758
1	1	1	0.917

The over-specified model showed similar results. Table 5.8 and Table 5.9 provide the attribute parameter estimates and the probability of each individual attribute mastery. The results

are not very different from those in the correct model; attribute *RU* is dependent upon attribute *APP* and *PI* because the probability of *RU* mastery increases as a respondent masters either *APP*, *PI*, or both. Again, the probability of *RU* mastery does not increase as a student masters attribute *MC*, suggesting that *RU* is possibly not depending solely on *MC* or a combination of *APP* and *MC*. When all three remaining attributes were mastered, the probability of *RU* mastery approached 1.00. Since the over-specified attribute structure included the pathway between *APP* and *PI*, with *APP* being the parent attribute, the probability of *PI* mastery increased from 0.26 (not depending on any parent attribute as shown in the correct model) to 0.67 (depending on *APP* as shown here in the over-specified model).

The question is if this over-specified structure is better, the answer is two-fold: (1) the overall model fitting indices point out that the deviance for the two models are comparable, with the DIC equaling 27462.36 for the correct model and 27420.05 for the over-specified model, and the over-specified model fitting slightly better (lower DIC); (2) with the fact that the attribute parameter and item parameter estimates converged equally well in terms of the percentage of R-hat values greater than 1.2, the parameter estimates in both models performed well (see Table 5.1). In short, the correct model and the over-specified model fit almost as well, which indicates that the estimates from both the correct model and the over-specified model are reliable.

Table 5.8

Mastery status of attributes for the over-specified model

	Logit	Probability of Mastery	Probability of Non-Mastery
APP	0.212	0.553	0.447
PI	−1.047	0.260	0.740
APP.PI	1.763	0.672	0.328
MC	0.985	0.728	0.272

Note. APP.PI = PI is dependent upon APP

Table 5.9

Probability of RU mastery for the over-specified model

APP	PI	MC	Probability of RU Mastery
0	0	0	0.206
0	0	1	0.123
0	1	0	0.402
1	0	0	0.305
0	1	1	0.440
1	0	1	0.267
1	1	0	0.750
1	1	1	0.916

In addition, Table 5.10 and Table 5.11 present the attribute parameter estimates and the probability of each individual attribute mastery for the under-specified model. The results are not very different compared with those from the correct model or the over-specified model: attribute *RU* is dependent upon the attributes *APP* and *PI* because the probability of *RU* mastery increases as a respondent masters either *APP*, *PI*, or both. Since the under-specified attribute structure removed the pathway between *RU* and *MC*, with *MC* being the parent attribute, the probability of *RU* mastery has decreased, meaning that the mastery of *RU* has been affected by this change. This indicates that the pathway from attribute *MC* to attribute *RU* is necessary.

However, as far as structure comparison is concerned, the overall model fit indices showed that the deviance for the three models - the correct model, the over-specified model, and the under-specified model- are comparable, yet the over-specified model still fits slightly better, followed by the under-specified model and the correct model (see Table 5.1). In addition, the attribute parameter and item parameter estimates in all three models converged well in terms of the percentage of R-hat values greater than 1.2 (see Table 5.1). In short, the correct model, the over-specified model, and the under-specified model fit almost as good, but the correct model yielded slightly more accurate student mastery classification.

Table 5.10

Mastery status of attributes for the under-specified model

	Logit	Probability of Mastery	Probability of Non-Mastery
APP	0.527	0.629	0.371
PI	−0.149	0.463	0.537
MC	0.640	0.655	0.345

Table 5.11

Probability of RU mastery for the under-specified model

APP	PI	MC	Probability of RU Mastery
0	0	0	0.184
0	1	0	0.454
1	0	0	0.280
1	1	0	0.861

Finally, attribute level results from the linear model are presented in Table 5.12. The results are very different compared with those from the other three model since this is a linear model. The probability of mastery of attribute *PI* increased from 0.28 to 0.84, the probability of mastery of attribute *MC* went from 0.29 to 0.97, and the probability of *RU* mastery increased from 0.25 to 0.71.

Table 5.12

Mastery status of attributes for the linear model

	Logit	Probability of Mastery	Probability of Non-Mastery
APP	0.242	0.560	0.440
APP.PI	1.670	0.842	0.158
PI	−0.937	0.282	0.718
PI.MC	1.801	0.970	0.030
MC	−0.874	0.294	0.706
MC.RU	1.984	0.709	0.291
RU	−1.095	0.251	0.749

Note. APP.PI = PI is dependent upon APP; PI.MC = MC is dependent upon PI; MC.RU = RU is dependent upon MC

In conclusion, the hypothesis related to the real data stating that there are hierarchies among the four attributes was plausible. As expected, by mastering one or more attribute, the probability of *RU* mastery increased quite significantly. In spite of the fact that the probability of *RU* mastery does not seem to increase significantly by mastery of attribute *MC* or the combination of *MC* and *APP*, the probability of *RU* mastery does increase evidently by mastery of the combination of *MC* and *PI*. Hence, we are convinced that the dependency assumption between attribute *RU* and the remaining three attributes holds.

Results for Mastery Classification

As for attribute classification, the individual attribute mastery proportion for each attribute structure is shown in Figure 5.8. Across the four models, *MC* is the attribute that was mastered by most respondents, followed by attribute *APP*, *PI* and *RU*. As shown in the figure, attribute *RU* was mastered by less than half of the respondents; the possible explanation is that attribute *RU* was always dependent upon at least one of the remaining three attributes.

Comparing mastery classification rates from the correct model with the over-specified model, it is interesting to note that the dependency between *PI* and *APP* did not increase the mastery classification rate of either attribute *PI* or attribute *RU*. The former was directly dependent upon the mastery of *APP*, and the latter was dependent on the other three attributes as a whole structure. This also suggests that the extra dependency between *PI* and *APP* does not benefit model fit, parameter estimation, or respondent's mastery classification. Thereby, the over-specified model does not perform better than the correct model.

As for the under-specified model, the pathway from attribute *MC* to attribute *RU* was removed. However, the dependency removed between *MC* and *RU* only lowered the mastery classification rate of attribute *RU* by 1%. Moreover, the classification of mastery to attribute *MC*

also dropped a bit, with the rate in the under-specified model slightly lower (94.97%) than in the other two models. This also indicated that the removed dependency between *RU* and *MC* does not significantly affect parameter estimation or respondents' mastery classification.

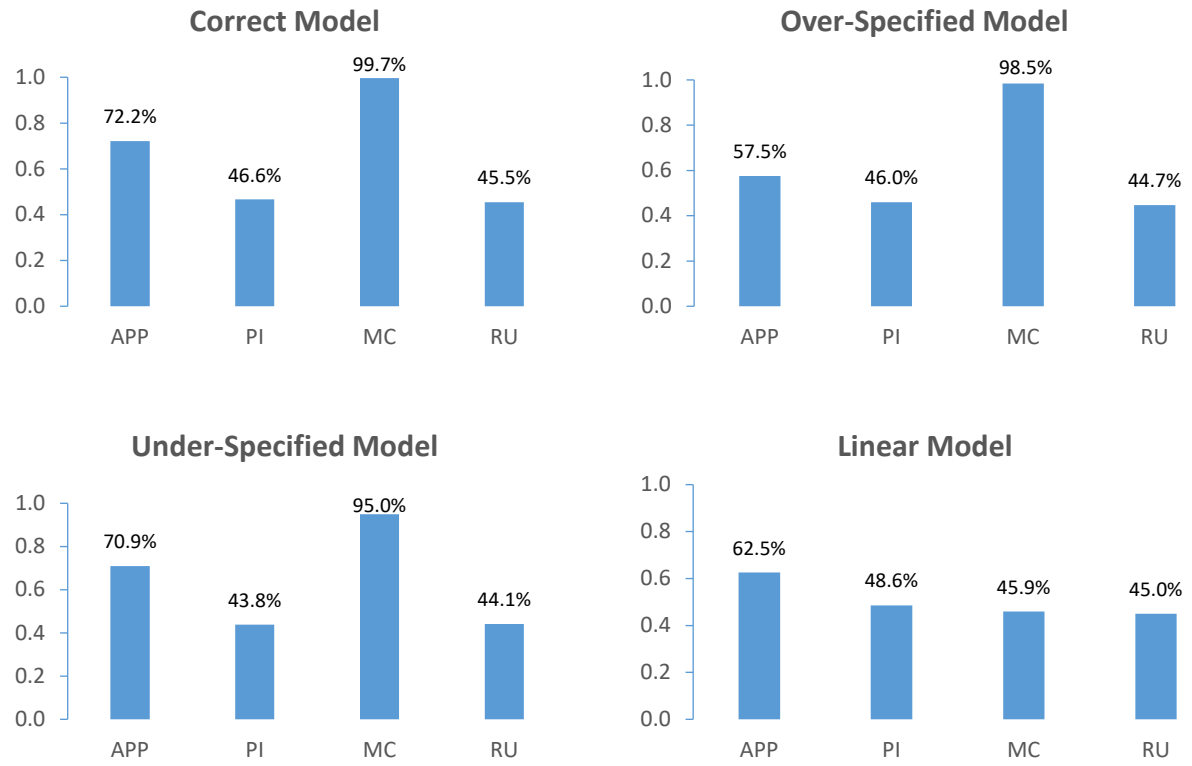

Figure 5.8. Proportion of Attribute Mastery by attribute structure. The figure displays the proportion of mastery (vertical axis) by attribute (horizontal axis).

Lastly, it was noted that mastery classification rates for attribute *APP*, *PI* and *RU* are quite comparable across four models: the correct model, the over-specified model, the under-specified model, and the linear model, whereas the mastery classification rate for attribute *MC* has significantly decreased from over 90% to 45.89% in the linear model. The dependency between *MC* and *PI* significantly impacted the mastery classification rate of attribute *MC* in the linear model. However, the mastery classification rate for attribute *RU* did not drop significantly because it is now only depending upon one attribute - attribute *MC*. This also points out that the sole dependency between *RU* and *MC* does not impact respondents' mastery classification

significantly. However, the model fit and parameter estimates indicated that the linear model has the largest DIC, which indicates a worst fit to the data. This conclusion is expected since we assumed that the linear model is a wrong model, given that the DTMR test was built under a multi-dimensional framework.

Although the classification of masters versus non-masters for each attribute does not vary much as the attribute structure changes, except for the linear model. Based on the model fit and model convergence indices, we believe that the parameter estimates and mastery classification from the correct model is more reliable than the other three models. Since the classification of mastery for attribute *PI* did not increase by adding *APP* as the parent to *PI*, the pathway between *APP* and *PI* is not tenable. Moreover, the classification for attribute *RU* did not drop or increase by removing attribute *MC* as the parent to *RU*, thus the pathway between *MC* and *RU* is not plausible either. However, the linear model presented a significant decrease in *MC* masters when treating *PI* as the parent to *MC*, though this finding needs to be confirmed in future research.

CHAPTER SIX

Discussion and Conclusion

This section provides a discussion of the findings produced in Chapters 4 and 5, answering the questions we raised in the methods section. To this point, this dissertation explored the impact of attribute structure, as well as the effect of sample size and test length, to parameter recovery and student mastery classification. The simulation models were inspired by the DTMR Fraction test data, as well as the varied conditions – sample size and test length – in each of the models. Each framework contained two levels: (a) the proficiency model and (b) the evidence model; thereby, each data set was generated and estimated using these two levels. With integrated LCDM and the Bayesian Network, the statistical tool estimated the parameters in the evidence and proficiency levels simultaneously across models. The performance of the models was compared at (1) model level – convergence, (2) the item level – item parameter recovery, (3) the attribute level – attribute parameter recovery, and (4) the person level – student mastery classification. This study manipulated three factors in the simulation study: (1) the sample size, (2) the test length – number of items per single attribute, and (3) the statistical models used for both data generation and estimation for examining the underlying attribute structure.

Impact of Attribute Structure

The differences in the estimation procedures produced differences in the parameter recovery and classification. Results indicated that the overall bias for both attribute intercept and item intercept are slightly smaller than those for main effect terms and interaction terms. The reason is that intercept by nature is the probability of mastering an attribute by chance or correctly answering an item without mastering the corresponding attribute. In IRT, this will be

the guessing parameter that does not depend on students' knowledge/ability level. With that in mind, the RMSE also showed smaller values for intercept than for main effect and interaction.

The parameters with parent parameters present were estimated more accurately. In our case, the main effect parameter exhibited the smallest bias values for attributes with over-specified pathways, and the item main effect had the least magnitude in bias. Although the RMSE values did not show a consistent pattern in terms of which parameter was estimated the most accurately, the item main effect was estimated better than item interaction parameters, indicating that items were estimated more accurately when they depended on corresponding attributes.

Moreover, the sample size determines the consistency among the estimates, and the test length is related to the accuracy of the estimates. As the sample size increases, the variation among estimates becomes smaller for items and attributes. When the test form is longer, the parameter estimates become more accurate, especially for more complex structures. Taken as a whole, the condition in which data sets were created and estimated using the Bayesian network with a large sample size and a long test length shows the best results. The student classification rate in the corresponding condition also yielded a result that the estimated classification rate is the lowest. One of the possible explanation is that DCM reliability is higher, and thereby does not require super long test length and large sample size. Further exploration of this issue should be interesting.

Most importantly, the simulation study compared the performance of four attribute structures: the hypothetically correct model, the over-specified model, the under-specified model, and the linear model, under the presence of attribute hierarchy. The data generated using the correct model and estimated in the linear model resulted in a reduction of classification

accuracy. The remaining three estimation models yielded comparable results in terms of classification accuracy for the test as a whole, with the under-specified model performing slightly better than the correct model and the over-specified model.

However, due to the diagnostic features of LCDM, we also care about the classification accuracy on each individual attribute in the test. For independent attributes, attributes that do not have parent attributes, the under-specified model provided slightly more accurate estimation on mastery classification, followed by the correct model and the over-specified model. In general, the specification of the attribute structure does not have much impact on the mastery classification of the whole assessment, yet it does have some effect in improving individual attribute mastery probability. Therefore, in a multidimensional test, depending on the purpose of the test and how the test is going to be used in real-world practice, the most appropriate attribute structure should be adapted.

In addition, the wrong model, the linear model, was included in this study as an illustration of how a linear attribute hierarchy could present the intentionally built multidimensional framework. It should be noted that the linear model estimating the multidimensional data does not yield a way-off model fit, and is a quite comparable model fit compared to other multidimensional models. However, the parameter estimation, as we discussed in the previous chapters, indicated more deviance from the true parameters. Overall, for a multidimensional test, LCDM is a preferred model in terms of testing the attribute structure and the standards of test item quality. Even if the linear structure is underlying the progressions of a learning map, the LCDM is still recommended, along with Bayesian networks.

In short, in a large scale assessment, the underlying map structure upon which the assessment based on is crucial for students' performance on skill level. Specifically, how well the

skills are structured in an assessment determines students' mastery classification on each of the skills as well as the assessment per se. Moreover, although the attribute structure does not influence students' performance on item level, items are written to certain skills, which means that how well the skills are specified, and how well the items are developed are crucial in an assessment. A set of well-delineated items in an assessment would improve the accuracy of students' mastery classification, and therefore leads to more accurate inference and decision making.

Methodological Contributions

In education and social science, many research studies assumed a hierarchical structure for attributes underlying a test, and one of the purposes of this study is to explore the presence of hierarchical structure for attributes mastered by respondents. Diagnostic classification models (DCMs) are well employed to classify students based on a series of discrete attributes. The Loglinear Cognitive Diagnosis Model (LCDM) is one of the statistical latent class-based models. In addition, the LCDM, within the DCM framework, assumed that all patterns of attributes are present, not specifically in a hierarchical sense. Hence, LCDM has been adapted as the statistical model to explore the potential attribute hierarchies.

The simulation studies together with the DTMR Fraction test data analysis provided evidence that the LCDM can be used to detect hierarchical attribute structures. Bayesian Networks using EM algorithm and MCMC estimation has also proven to be a good approach integrated with LCDM for attribute structure detection. More specifically, the simulation studies provided evidence that the LCDM can be used to detect hierarchical attribute structures. The LCDM integrated with Bayesian networks provided an empirical method to assess the presence or absence of an attribute hierarchy compared with other models within the DCM framework,

such as the DINA and DINO models, which are also commonly used DCMs under the presence or absence of attribute hierarchies. DINA and DINO models were subsumed by the LCDM with a simplified structure, and proven for low ability in attribute hierarchy detection (Templin & Bradshaw, 2014). On the other hand, a nested model within LCDM, known as Hierarchical classification diagnostic model (HCDM), presumes the hierarchies among the attributes, and was proven to be a method of detecting attribute hierarchies. However, compared to HCDM, LCDM allows more freedom in attribute structure in terms of attribute hierarchies, as discussed previously.

Although the simulated and real data results in this study compared different attribute structures in order to acknowledge statistical model-data fit considerations and classification accuracy, it is critical to emphasize that the selection of the psychometric model should be determined with respect to the purpose of the test and the attribute structure should be decided by the blueprint of the test development and test purpose. Although the LCDM along with a linear attribute hierarchy may be seen as an approximation of a 3-PL model, the models serve different purposes. The DCMs (to which the LCDM belongs) classify respondents as either masters or non-masters of a test, whereas IRT models scale respondents. In short, DCMs serve as a means to directly classify students as opposed to putting them on a scale. Thereby, DCMs are more appropriate for tests seeking to measure more than one latent attribute. To this extent, multidimensional IRT could serve the same purpose, yet the model requires longer test forms in order to precisely scale respondents to multiple latent attributes (Templin & Bradshaw, 2013). All in all, LCDM, under the DCM framework, integrated with Bayesian Networks is a practical choice for multidimensional test scoring.

Conclusion

Attribute hierarchies, if present, are important structural features that provide practical suggestions about the model LCDM and Bayesian Networks, as well as the attributes, to both researchers and test developers. In practice, if attribute hierarchies are present, instructional plans could be made for target students to acquire the non-mastered skills or knowledge, as an alternate approach to the target skill. That being said, if the attribute hierarchy does not exist in an attribute structure, such potential benefits would not hold.

The results of this study supported many of the findings from prior research with respect to the models for attribute hierarchy detection. In short, the LCDM emerges as a psychometric model that can be used to detect attribute hierarchies and model attributes using the full set of parameters from the fully crossed model. In addition, Bayesian Networks are a well-known statistical method that estimates parameters in a machine-leaning environment. Therefore, LCDM could be added into the current methodological toolbox for researchers and practitioners with its modeling options. Furthermore, as general multidimensional test construction becomes more common in practice, we anticipate better applications of the LCDM for the classification of students according to theorized attribute hierarchies.

This study is limited in that the number of attributes included in the models are not big, which leads to less complex attribute structures. Because the attribute structures were inspired by the DTMR Fraction test data, the overall hierarchies in the attribute structures were pre-determined, leaving other possibilities out. Future research could endeavor to investigate more complex attribute structures and compare different models in detecting attribute hierarchies.

References

Anderson, J. O. (1999). Does complex analysis (IRT) pay any dividends in achievement testing? *The Alberta Journal of Educational Research, 45(4),* 344-352.

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

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics*, *7*(4), 434-455.

Cohen, J. (1992). A power primer. Psychological bulletin, 112(1), 155.

Choi, H. J., Templin, J. L., Cohen, A. S., & Atwood, C. H. (2010). The impact of model misspecification on estimation accuracy in diagnostic classification models. In *meeting of the National Council on Measurement in Education (NCME), Denver, CO*.

De La Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of educational and behavioral statistics*, *34*(1), 115-130.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). Bayesian data analysis, 2nd edn. Texts in Statistical Science.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. *In Bayesian Statistics 4, Bernardo, J. M., Berger, J. O., Dawid, A. P. and Smith, A. F. M. (eds.),* 169-193. Oxford: Oxford University Press.

Gierl, M. J. (2007). Making diagnostic inferences about cognitive attributes using the rule-space model and attribute hierarchy method. Journal of Educational Measurement, 44, 325-340. doi: 10.1111/j.1745-3984.2007.00042.x

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.

Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality* (Doctoral dissertation, ProQuest Information & Learning).

Heckerman, D., Geiger, D., & Chickering, D. M. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine learning*, *20*(3), 197-243.

Heidelberger, P., & Welch, P. D. (1981). A spectral method for confidence interval generation and run length control in simulations. *Communications of the ACM*, *24*(4), 233-245.

Heidelberger, P., & Welch, P. D. (1983). Simulation run length control in the presence of an initial transient. *Operations Research*, *31*(6), 1109-1144.

Huff, K., & Goodman, D. P. (2007). The demand for cognitive diagnostic assessment. In J. P. Leighton & M. Gierl (Eds.), Cognitive diagnostic assessment for education: Theory and applications (pp. 19-60).

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.Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Routledge.

Lord, F. M. (2012). *Applications of item response theory to practical testing problems*. Routledge.

Maris, E. (1995). Psychometric latent response models. *Psychometrika*, *60*(4), 523-547.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational researcher*, *23*(2), 13-23.

Miles, J. & Shevlin, M. (2001) Applying Regression and Correlation: A Guide for Students and Researchers. Sage: London.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *RESEARCH REPORT-EDUCATIONAL TESTING SERVICE PRINCETON RR, 16*.

Nielsen, T. D., & Jensen, F. V. (2009). *Bayesian networks and decision graphs*. Springer Science & Business Media.Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann.

Pearl, J. (1998). Graphs, causality, and structural equation models. *Sociological Methods & Research*, *27*(2), 226-284.

Plummer M (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. URL http://citeseer.ist.psu.edu/plummer03jags.html.

Plummer, M. (2012). JAGS: Just Another Gibbs Sampler v. 3.3. 0.

Plummer, M. (2013). rjags: Bayesian graphical models using MCMC. *R package version*, *3*(10).

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

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

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 263-331). New York: American Council on Education/Macmillan.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64*(4), 583-639.

Stone. C. A., & Hansen, M. A. (2000). The effect of errors in estimating ability on goodness-of fit tests for IRT models. *Educational and Psychological Measurement, 60,* 974-991.

Su, Y. S., & Yajima, M. (2012). R2jags: A Package for Running jags from R. *R package version 0.03-08, URL http://CRAN. R-project. org/package= R2jags*.

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

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

Trout, D. L., & Hyde, B. (2006, April). Developing score reports for statewide assessments that are valued and used: Feedback from K-12 stakeholders. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

von Davier, M. (2005). A general diagnostic model applied to language testing data. ETS Research Report Series, 2005(2), i-35.

Yang, X., & Embretson, S. E. (2007). Construct validity and cognitive diagnostic assessment. In J. P. Leighton & M. Gierl (Eds.), Cognitive diagnostic assessment for education (pp. 119-145). New York: Cambridge University Press.

Zwaan, R.A. & Singer, M. (2003). Text Comprehension. In A.C. Graesser, M.A. Gernsbacher & S.R. Goldman (Eds.), Handbook of Discourse Processes. Mahwah, NJ: Lawrence Earlbaum.

Appendix A: JAGS Code for Data Estimation

JAGS code for data estimation

```
# N: the number of students
# K: the number of attributes
# J: the number of items
# alpha: student mastery profile

bayes.mod<-function(){
  #proficiency model

  for (i in 1:N){
    for (k in 1:3){

      logit(pi[i,k]) <- lambda0[k]    #app pi, mc, do not have parent node
      alpha[i,k]~dbern(pi[i,k])
    }

    for (k in 4:4){ #RU

      logit(pi2[i,k])<-lambda0[k]+lambda1.main[1]*alpha[i,1]+    #ru.app
        lambda1.main[2]*alpha[i,2]+                              #ru.pi
        lambda1.main[3]*alpha[i,3]+                              #ru.mc
        lambda1.int[1]*alpha[i,1]*alpha[i,2]+                    #ru.app.pi
        lambda1.int[2]*alpha[i,1]*alpha[i,3]+                    #ru.app.mc
        lambda1.int[3]*alpha[i,2]*alpha[i,3]+                    #ru.pi.mc
        lambda1.int[k]*alpha[i,1]*alpha[i,2]*alpha[i,3]          #ru.app.pi.mc

      alpha[i,k]~dbern(pi2[i,k])
    }
  }
  # PRIORS
  for (k in 1:4){
    lambda0[k]~dnorm(0,10)
    lambda1.int[k]~dnorm(0,10)
  }
  for (k in 1:3){
    lambda1.main[k]~dnorm(0,10)
  }

  #evidence model
  ### for items measuring only one attribute ###
  ### APP
  for (i in 1:N){
    for (j in 1:J_1){
      logit(pr_app[i,j])<-intercept.app.item[j]+main.app.item[j]*alpha[i,1]
      score_1[i,j]~dbern(pr_app[i,j])
    }
```

```
 ### PI
 for (j in 1:J_2){
  logit(pr_pi[i,j])<-intercept.pi.item[j]+main.pi.item[j]*alpha[i,2]
  score_2[i,j]~dbern(pr_pi[i,j])
 }

 ### MC
 for (j in 1:J_3){

  logit(pr_mc[i,j])<- intercept.mc.item[j]+main.mc.item[j]*alpha[i,3]
  score_3[i,j]~dbern(pr_mc[i,j])
 }

 ### RU
 for (j in 1:J_4){
  logit(pr_ru[i,j])<-intercept.ru.item[j]+main.ru.item[j]*alpha[i,4]
  score_4[i,j]~dbern(pr_ru[i,j])
 }

 ### items measuring two attributes ###
 ### PI-MC
 for (j in 1:J2){
  logit(pr2[i,j])<-intercept.pi.mc.item[j]+gamma1.pi[j]*alpha[i,2]+
    gamma1.mc[j]*alpha[i,3]+
    gamma1.pi.mc[j]*alpha[i,2]*alpha[i,3]

  score2[i,j]~dbern(pr2[i,j])

 }
    ### RU-MC
 for (j in 1:J3){
  logit(pr3[i,j])<-intercept.ru.mc.item[j]+gamma1.ru[j]*alpha[i,4]+
    gamma1.mc2[j]*alpha[i,3]+
    gamma1.ru.mc[j]*alpha[i,4]*alpha[i,3]
  score3[i,j]~dbern(pr3[i,j])
 }
    ### RU-PI
 for (j in 1:J4){
  logit(pr4[i,j])<-intercept.ru.pi.item[j]+gamma1.pi2[j]*alpha[i,2]+
    gamma1.ru2[j]*alpha[i,4]+
    gamma1.ru.pi[j]*alpha[i,2]*alpha[i,4]
  score4[i,j]~dbern(pr4[i,j])

 }
}
```

```
  # priors
 for (j in 1:J_1){
  intercept.app.item[j]~dnorm(0,10)
  main.app.item[j]~dnorm(0,10)%_%T(0,)
 }
 for (j in 1:J_2){
  intercept.pi.item[j]~dnorm(0,10)
  main.pi.item[j]~dnorm(0,10)%_%T(0,)
 }
 for (j in 1:J_3){
  intercept.mc.item[j]~dnorm(0,10)
  main.mc.item[j]~dnorm(0,10)%_%T(0,)
 }
 for (j in 1:J_4){
  intercept.ru.item[j]~dnorm(0,10)
  main.ru.item[j]~dnorm(0,10)%_%T(0,)
 }

 for (j in 1:2){
  intercept.pi.mc.item[j]~dnorm(0,10)
  gamma1.pi[j]~dnorm(0,10)%_%T(0,)
  gamma1.mc[j]~dnorm(0,10)%_%T(0,)
  gamma1.pi.mc[j]~dnorm(0,10)%_%T(0,)

  intercept.ru.mc.item[j]~dnorm(0,10)
  gamma1.ru[j]~dnorm(0,10)%_%T(0,)
  gamma1.mc2[j]~dnorm(0,10)%_%T(0,)
  gamma1.ru.mc[j]~dnorm(0,10)%_%T(0,)
 }
 for (j in 1:4){
  intercept.ru.pi.item[j]~dnorm(0,10)
  gamma1.pi2[j]~dnorm(0,10)%_%T(0,)
  gamma1.ru2[j]~dnorm(0,10)%_%T(0,)
  gamma1.ru.pi[j]~dnorm(0,10)%_%T(0,)
 }
} #end of model
```