University of Massachusetts Amherst

ScholarWorks@UMass Amherst

Doctoral Dissertations

Dissertations and Theses

March 2019

# Bringing Learning Back In: Examining Three Psychometric Models for Evaluating Learning Progression Theories

Duy Pham

# Bringing Learning Back In: Examining Three Psychometric Models for Evaluating Learning Progression Theories

A Dissertation Presented

by

DUY NGOC PHAM

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

February 2019

Research in Educational Measurement and Psychometrics

College of Education

Bringing Learning Back in: Examining Three Psychometric Models for Evaluating Learning Progression Theories

A Dissertation Presented

by

DUY NGOC PHAM

Approved as to style and content by:

Craig S. Wells, Chair

Ronald K. Hambleton, Member

Stephen G. Sireci, Member

Malcolm I. Bauer, Member

Jennifer Randall
Associate Dean of Academic Affairs
College of Education

# DEDICATION

*“To all great teachers I have had throughout the course of my life.”*

There are many things I have yet to know, but one thing I am certain about is that I could not have been here today writing these words without the support and inspirations I have received from great teachers, including my parents who not only gave me my physical body but also nurtured my spirit and determination. My mom was my first Math teacher and she taught me, through her own life, what love and optimism can do to my family and the schools that she worked for. My dad taught me the importance of learning and the beauty of persistence through his amazing life as an orphan. In middle school, I had classes with a great teacher, Mr. Pham Viet Hung, who planted the seed of thinking about how to improve education of Vietnam in my mind so much so that it has become my career and dedication for life. Later, I met my wife. And, I also wish to dedicate this dissertation to her since she, through her purity and patience, guided me toward the heavenly territory of Buddhism. Without relying on the teaching of and spiritual belief in Buddha, hardly could I have survived the Ph.D. program and dissertation process.

As for perspectives acquired in the West, I was so fortunate to have Philip Altbach from Boston College as my first academic advisor when I started to learn about educational systems worldwide. Not only has his career been so impressive, his caring mindset and thoughtfulness have also been very inspirational to me. This dedication would be half-baked if I neglected to mention Ronald Hambleton, Stephen Sireci and Craig Wells from UMass Amherst. Ron is the most productive professor who I have ever had a chance to work with. After working with him for at least four years, I am still impressed with his research and teaching passion. He and I used to start every one-to-one meeting by talking about his wonderful wife, Else, and my family. Similarly with Steve, he asked me on my first week at UMass if I needed to borrow him some money to bring my family to the U.S. He is the first one who comes to mind when I need to think about validity. He is nothing but warm-hearted, smart and thoughtful. In addition, Craig Wells is a unique professor for whom I will be forever grateful. To me, he is not a man of words, he is a man of thoughts and profound ideas. Craig understands his students and gives them the freedom to pursue their research interests and professional dreams.

Finally, I would like to recognize the last two great teachers that I have had more recently in my family. Surprisingly and thankfully, my sons seem to have taught me more than I have taught them. Indeed, they showed me how to have fun and to be happy with tiny little things such as running down a lawn covered with lots of leaves in the Fall or shoveling the snow during the freezing winter of New England. They teach me to love and enjoy but not to hate and complain. To me, they are a lively illustration of the organic interaction between living and learning. Through their own early lives, they have provided me these valuable lessons at no cost. I can’t think of anything sweeter than working on a dissertation of learning theories while you have two little ones running around and showing you how they learn new knowledge and sharpen their skills day in and day out. With that said, I truly believe they have saved my dissertation.

# ACKNOWLEDGMENTS

No one could finish a dissertation alone and I am not an exception. I received support from many organizations and individuals before and during my doctoral program. By no way could I list them all out here. However, given the limited space, I would like to express my thankfulness to my home country of Vietnam, three organizations and a few individuals for being alongside with me. My country immerged me naturally into a problem of how to improve an educational system. Sometimes, I wish the problem didn’t even exist. Yet, it is there so visible for Vietnam and many other countries that it has intrigued me and become the end purpose that helps me navigate through this life.

In what follows, I would like to mention the three organizations that I feel deeply thankful for. First, Educational Testing Service has been so generous to allow me to access their learning progression materials and data. In my mind, their learning progression team had done all the hard work to define theories, develop assessment items and collect student responses. Then, they were more than willing to pass the asset along to us at UMass to help them analyze the data. To me, that was one of the most generous research collaborations across organizations that I have ever seen. Second, my thanks also go to the Center of Educational Assessment at UMass. Throughout my program, the center supported me and my family financially, academically and spiritually. As I live far from home, the center and students affiliated with it have truly become our extended families. Last, Curriculum Associates warmly welcomed me as a summer intern and kept supporting me financially while I am wrapping up my dissertation. How the company has treated me means a great deal to me as an international student with family living in the United States of America at this point in time.

For individuals, I do not have enough space here to mention all the people who I would like to thank for being part of my life, supporting and trusting me while I was working on my doctoral program. Nonetheless, there are some special people that I wish to bring their names to light here. First, Malcolm Bauer from ETS inspired me by his works on learning progressions. It was so kind of him to support me throughout the empirical data analysis. Also, I was so fortunate to have him serve in my dissertation committee. Second, it is my privilege to have Ronald Hambleton and Stephen Sireci in my dissertation committee. Definitely, Ron opened up my mind to the line of research on learning hierarchy that he was a part of all the way in the 1960s and 1970s. He paid attention to detail and helped me improve the clarity and concision of my writing. As for Steve, he helped me hold the concept of validity in mind while I was exploring the models and the theories of learning progressions. Lastly, I don’t think I have enough words to express my thankfulness to Craig Wells. His patience, trust and research method insight that he offered me were nothing but incredible and inspirational.

Before going into the detail of this study, let me share the story behind the title of this dissertation. A few years ago, I told a former professor of educational history and change of mine from Boston College that I was working on learning progressions and asked him for a few books to read to help me figure out how to bring this idea closer to the classroom. He responded to me with encouragement and a few references. One of them was a book entitled “Bringing knowledge back in: From social constructivism to social realism in the sociology of education” by Michael F. D. Young. I am still reading the book and  haven’t finished it yet, but I really like the title. That is why I changed the word “knowledge” into “learning” to create the title of my dissertation. In doing that, I am well aware that the title is more of an ideal scenario that will take time and resources to be realized in every educational system. The subtitle of this dissertation seems to reflect the scope and purpose of this study much better than the main title. Currently, I view the investigation of model effectiveness in evaluating learning progressions is one of the many steps to bring learning and knowledge back in to schools and educational activities.

Finally, this study used data collected from a project funded by support from the Institute of Educational Sciences, U.S. Department of Education, through Grant Reference R305A100518 and is partially supported by Educational Testing Service. This support is highly appreciated and gratefully acknowledged. Any opinions, findings and conclusions expressed in this dissertation are those of the author and do not necessarily reflect the views of the Institute of Educational Sciences and Educational Testing Service.

# ABSTRACT

## BRINGING LEARNING BACK IN: EXAMINING THREE PSYCHOMETRIC MODELS FOR EVALUATING LEARNING PROGRESSION THEORIES

FEBRUARY 2019

DUY NGOC PHAM, B.S., HANOI NATIONAL UNIVERSITY OF EDUCATION
M.S., PARIS-SUD UNIVERSITY
M.A., BOSTON COLLEGE

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Dr. Craig S. Wells

Learning progressions provide potentially valuable information to teachers about how to develop a scope and sequence for a group of learning objectives. However, for the learning progressions to be valuable, the progressions must be supported. Although there are several approaches and models that can be used to evaluate the validity of a learning progression, there is a dearth of research examining the advantages and limitations of each approach. The purpose of this study was to examine a multi-dimensional IRT model and two cognitive diagnostic models (DINA and HO-DINA) for evaluating two learning progressions via a simulation study. In addition, the models were applied to empirical data to determine if the models provided consistent results. The results from the investigation indicated that five methods of using the model and statistical methods derived from them to testify learning level order could complement each other. None of the methods worked dominantly better than the others but they all deemed useful in certain contexts. With respect to assessing the possible links among levels across progressions, the degree to which the model recovered the true information in the simulation studies varied depending on the model and the magnitude of the difference between the learning levels. The more distant the levels were, the more accurate the model became at recovering the true classification. For the empirical analysis, three models provided convergent evidence to support almost all the aspects of the theory underlying two progressions considered in this study. Statistical results also suggested a few revisions to make the theory more in line with the empirical evidence. Four limitations were discussed, and six future directions were elaborated to address the drawbacks of this study. Finally, three practical implications were presented as take-away messages from this dissertation.

## TABLE OF CONTENTS

Page

ACKNOWLEDGMENTS ........ vi
ABSTRACT ........ ix
LIST OF TABLES ........ xiii
LIST OF FIGURES ........ xv
CHAPTER
I. INTRODUCTION ........ 1
1.1. Educational Context of Study ........ 1
1.2. Purpose of Study ........ 4
II. LITERATURE REVIEW ........ 7
2.1. Overview of Learning Progression Theories ........ 7
2.1.1. The “Forgotten” Concept of Learning Hierarchy ........ 8
2.1.2. The Concept of Learning Progressions ........ 10
2.1.3. An Example of a Learning Progression Theory ........ 15
2.1.4. Validating Learning Progression Theories ........ 19
2.2. Psychometric Models to Evaluate Learning Progression Theories ........ 22
2.2.1. Item Response Theory ........ 22
2.2.2. Cognitive Diagnosis Models ........ 28
2.3. The Relationship between IRT and CDMs ........ 34
2.4. Summary of the Psychometric Models ........ 37
2.5. Tables and Figures for Chapter 2 ........ 43
III. METHOD ........ 46
3.1. Study 1: MIRT-SS as the Generating Model ........ 47
3.1.1. Data Generation ........ 47
3.1.2. Parameter Estimation ........ 49
3.1.3. Data Analysis ........ 51
3.2. Study 2: HO-DINA as the Generating Model ........ 56
3.2.1. Data Generation ........ 57
3.2.2. Parameter Estimation and Data Analysis ........ 60
3.3. Study 3: An Empirical Application ........ 61
3.3.1. Data ........ 62
3.3.2. Parameter Estimation and Model-data Fit ........ 63
3.3.3. Data Analysis ........ 64
3.4. Summary of Research Method ........ 65
3.5. Tables and Figures for Chapter 3 ........ 67
IV. RESULTS ........ 68
4.1 Study 1: MIRT-SS as the Generating Model ........ 68
4.1.1. Claim 1: Ordering of Learning Levels within Each Progression ........ 69
4.1.2. Claim 2: Co-occurrence of Learning Levels across Progressions ........ 71
4.2. Study 2: HO-DINA as the Generating Model ........ 76
4.2.1. Claim 1: Ordering of Learning Levels within Each Progression ........ 77
4.2.2. Claim 2: Co-occurrence of Learning Levels across Progressions ........ 79
4.3. Study 3: An Empirical Application ........ 84
4.3.1. Calibration Results and Model-data Fit ........ 85
4.3.2. Claim 1: Ordering of Learning Levels within Each Progression ........ 90

4.3.3. Claim 2: Co-occurrence of Learning Levels across Progressions ....... 95
4.4. Chapter Summary ........................................................................... 97
4.5. Tables and Figures for Chapter 4 ................................................. 100
V. DISCUSSION .......................................................................................... 122
5.1. Simulation Studies ............................................................................ 122
5.1.1. Claim 1: Level Order ........................................................... 123
5.1.2. Claim 2: Level Link .............................................................. 124
5.1.3. Results across Models and Conditions ................................ 125
5.2. Empirical Study ................................................................................ 128
5.2.1. Claim 1: Level Order ........................................................... 128
5.2.2. Claim 2: Level Link .............................................................. 129
5.3. Conclusion ........................................................................................ 130
5.3.1. Limitations ........................................................................... 133
5.3.2. Future Directions ................................................................. 135
5.3.3. Practical Implications .......................................................... 138
APPENDIX: PERCENTAGES OF STUDENTS IN EACH PROFILE .............................. 142
REFERENCES ............................................................................................ 144

## LIST OF TABLES

Table Page

4.1.2. True Proportions of Students in Level Combinations ................................................101

4.1.3. Proportions of Students by Level Combination for MIRT ........................................101

4.1.4. Proportions of Students in Each Level Combinations by CDMs ................................102

4.1.5. Proportions of Students in Inconsistent Combinations................................................102

4.1.6. Classification Accuracy for True Conditions .............................................................103

4.1.7. Cross-model Classification Consistency for True Conditions....................................103

4.2.1. True and False Positive Rates for Extreme Difference Cases .....................................104

4.2.2. True and False Positive Rates for Moderate Difference Cases...................................105

4.2.3. Classification Accuracy and Consistency for Extreme Difference Cases ...................105

4.2.4. Proportions of Students in Level Links (Extreme Difference Cases)..........................106

4.2.5. Proportions of Students in Inconsistent Links (Extreme Difference Cases)...............107

4.2.6. Accuracy and Consistency for Moderate Difference Cases (Significant Cases).........108

4.2.7. Accuracy and Consistency for Moderate Difference Cases (Ordered Cases) .............108

4.2.8. Proportions of Students in Level Links (Moderate Difference Cases) ........................109

4.2.8. Proportions of Students in Inconsistent Links (Moderate Difference Cases).............110

4.3.1. Limited-Information Fit Statistics and Indexes for MIRT-SS....................................111

4.3.2. Summary Chen-Thissen LD X2 Fit Statistics for MIRT-SS ......................................111

4.3.3. Relative Fit Statistics for Fitting CDMs to Empirical Data........................................111

4.3.4. Absolute Fit Statistics for Fitting HO-DINA to Empirical Data ................................112

4.3.5. Absolute Fit Statistics for Fitting DINA to Empirical Data .......................................112

4.3.6. Summary of Item Parameter Estimates by Fitting CDMs to Empirical Data.............113

4.3.7. Summary of Discrimination Indexes of CDMs for Empirical Data ...........................113

4.3.8. Results of Two-sample T-tests Comparing Item Difficulties ......................................114

4.3.9. Attribute Locations for Six Data Sets ............................................................................114

4.3.10. Proportions of Students in the Learning Profiles by CDMs ......................................114

4.3.11. Proportions of Students in Level Combinations (LFPR1223)...................................115

4.3.12. Proportions of Students in Level Combinations (LFPR2334)...................................115

4.3.13. Proportions of Students in Level Combinations (LFPR3445)...................................116

4.3.14. Decision Consistency Between Pairs of Models ......................................................116

# LIST OF FIGURES

Figure Page

2.2.1. Path diagram of the MIRT-SS ..........................................................................45

4.1.1. True Positive Rates by Five Methods for True Scenario............................................117

4.1.2. False Positive Rates by Five Methods for False Scenario ..........................................117

4.1.3. Proportions of Students in Nine Level Links.............................................................118

4.2.1. Proportions of Students in Level Links (Extreme Difference Cases)..........................118

4.2.2. Proportions of Students in Level Links (moderate difference cases) ..........................119

4.3.1. Ordering of Difficulty Estimates of Items Measuring Different Levels......................120

4.3.2. Observed Proportions of Students in Level Combinations.........................................121

# CHAPTER I

# INTRODUCTION

## 1.1. Educational Context of Study

Learning, as a process of students acquiring and advancing knowledge and skills, has been at the heart of educational activities throughout modern human history (Faure et al., 1972). This view suggests that educators should prioritize resources to support student learning. If this vision is worth pursuing, educational assessments should play a critical role in sustaining and strengthening the learning process of students. The reason is that assessments, when they are properly aligned with curriculum and instruction, can facilitate student learning (Martone & Sireci, 2009).  To make assessments more directly useful for instruction and learning, instruments based on cognitive models and theories of learning are a possible solution (Kane & Bejar, 2014; Heritage, 2008). In short, the demand to assess learning that can support student academic growth and teachers to improve their instruction is an important part of education.

To meet this demand, assessments based on learning progressions have recently been proposed as a promising solution to bridge assessment information to student learning during instructional cycles. Major testing organizations have conducted studies and/or implemented developmental projects on learning progressions to build formative assessments that can capture student learning and measure student growth (Arieli-Attali, Wylie, & Bauer, 2012; Camara, O'Connor, Mattern, & Hanson, 2015). In general, learning progressions can be defined as empirically grounded and testable hypotheses of how the knowledge and skills of students develop and reach more sophisticated levels overtime with suitable instruction (Corcoran, Mosher, & Rogat, 2009). If multiple progressions are involved, we can define a theory of related learning progressions as descriptions and hypotheses that describe (i) how student's knowledge and skills develop, strengthen and advance from novice to mastery within the content of each progression, and (ii) the relationship of the learning process across the progressions.

If an educator obtains a supported theory of learning progressions, it can be useful in several perspectives. First, capitalizing on the concept of learning progressions, she/he could construct assessment systems that might produce meaningful information regarding student learning. For example, such a system would allow us to build measures that provide more reliable and valid inferences regarding student growth (Briggs, Diaz-Bilello, Peck, Alzen, Chattergoon, & Johnson, 2015; Thissen, 2015). Indeed, Briggs and Peck (2015) suggested the use of learning progressions to construct a vertical scale that includes two sub-scales. The first scale reflects the overall student achievement on a whole domain of content of a given grade, and the other scale measures student growth in regard to a learning progression within the domain.

Second, learning progressions and assessments based on the concept of learning progressions can be useful for instructional purposes (Daro, Mosher, & Corcoran, 2011). If such assessments are available, teachers can use them to obtain timely and reliable information of the learning status of each student in each point in time. This information is useful for them to provide feedback to students in regard to their learning and possible misunderstandings of the knowledge and skills defined by learning progression theories. The assessment results can also inform teachers to design or customize their instruction to meet the need of individual students. Even if the assessment might not be available, teachers can also use the theory as a reference point to set the right conditions for learning to foster a student’s deeper understanding of the content areas encapsulated by the progressions.

Third, learning progressions can be described as embedded within the scope of popular K-12 curriculua. For instance, Confrey, Maloney and Corley (2014) identified and elaborated 18 learning progressions within the Common Core State Standards for mathematics for the first nine grades (K-8). They then built an online platform to scaffold the curricula based on the learning progressions to support student learning and instruction (Confrey, Gianopulus, McGowan, Shah, & Belcher, 2017). In summary, learning progressions and assessments that capitalize on the concept of learning progressions can lead to meaningful applications in the real world of education to assist students to learn better and instructors to facilitate students to learn more effectively.

Regardless of the promising scenario described above that learning progressions could bring about, the realization of the idea to build learning progression-based assessments faces significant challenges (Briggs & Peck, 2015; Confrey, Jones, & Gianopulos, 2015). For instance, the construction of assessments using learning progressions might consume significant resources. More importantly, it requires that one would have been able to empirically validate the underlying theory of learning progressions before we can rely on the theory to build assessments that measure student learning and growth. In other words, the usefulness of assessments built on learning progressions depends, in large part, on the validity of the underlying theory of the progressions and the psychometric foundation to scale the assessment data. If the theory is not supported empirically, theory-informed inferences about student learning made from assessment results may not be valid, which means that it may not be useful for instructional purposes.

From a validity perspective, validating a learning progression framework requires collecting different sources of evidence to support or refute the claims postulated by the theory. For example, if the items were developed to identify the relative position of student learning in a progression, the response data should reveal empirical evidence that supports the correct identification of student learning levels. Similarly, if the theory predicted that a student in a learning stage in a progression can master a certain level of skills and knowledge of a different but related progression, the observed data should support this claim. To obtain the evidence to examine the claims, some statistical models can be useful to help draw some conclusions from learning progression data.

### 1.2. Purpose of Study

There are at least three families of psychometric models that have been used to analyze learning progression data (Pham, Bauer, Wylie, & Wells, 2017). The first one is based on a classical test theory (CTT) framework. The second one is based on modern test theory, or item response theory (IRT) models. During the last decade or so, cognitive diagnosis models (CDMs) have been adopted to analyze learning progression data (e.g., Chen, Zhang, Guo, Xin, 2017; Kizil, 2015; Pham et al., 2017). Traditionally, IRT and CDMs rely on different assumptions of the underlying latent variables. In an IRT framework, one assumes that the latent variables are continuous. Whereas, in CDMs, students are classified into a finite number of discrete latent profiles defined by the set of attributes measured by the assessment. However, de la Torre and Douglas (2004) proposed a higher-order cognitive diagnosis model (CDM) framework that assumes there are continuous latent variables, as in the case of IRT, that derive the joint distribution of the cognitive attributes. If the variable is unidimensional, the higher-order model can locate the attributes under the CDM framework in an increasing order. This feature of the model seems to be relevant to evaluate learning progressions.

Interestingly, some recent studies have fit both IRT and CDMs to the same data set of learning progressions (e.g., Chen et al., 2017; Kizil, 2015). In some cases, it was observed that both IRT and CDMs provide adequate model-data fit (Haertel, 1990). Under certain modeling settings, Haertel (1990) stated that some CDMs can be considered as special cases of IRT models. However, to date, there are no simulation studies that have evaluated the usefulness of these models in the context of learning progressions, especially for newer CDMs such as the higher-order models. Given the scarcity of literature on this topic, simulation studies that investigates the effectiveness of the models to evaluate learning progressions are useful to guide practices and suggest future directions. In this context, this study is an effort to shed light on the effectiveness of one IRT model and two CDMs in analyzing learning progression data under various practical conditions. The first model is the two-parameter logistic multidimensional IRT with simple structure (MIRT-SS). The two CDMs are (i) *deterministic input, noisy "and" gate (DINA),* (ii) and its higher-order version (HO-DINA or HO) (de la Torre & Douglas, 2004; Haertel, 1990). Two simulation studies will be conducted to investigate the effectiveness of the models when the true model and information about the learning levels is known. Then, an empirical study fitting the three models to response data collected from an assessment system to validate the theory we investigated in our previous works will be implemented.

The effectiveness of those models in evaluating such theories entails several aspects that will be examined in detail. In the first place, this study will shed light on how the IRT and CDMs are effective in recovering the ordering of learning levels in simulated conditions. In the second place, the effectiveness of the models will be investigated in regard to the second claim about the relationship between levels across learning progressions. Findings of the simulations are expected to inform the interpretation of results obtained from fitting the models to the empirical data. The empirical results will also be connected to prior validity evidence to draw conclusions about how effective the models are in analyzing the data.

The significance of this study can be visualized in two perspectives. The primary potential contribution of this study is that it can illuminate the comparative strengths and weaknesses of each model in analyzing the learning progression data in several realistic conditions and empirical data. In the second place, the analysis that fits the models to the empirical data would provide constructive information to the proposers of the theory to revise and improve their underlying theory.

The dissertation will be organized as follows. The next chapter, chapter 2, provides a comprehensive review of the literature that related to the purpose of this study. Then, a complete description of the method and three sub-studies carried out to shed more light on the effectiveness of the selected models in evaluating learning progressions will be presented in chapter 3. Two simulation studies, one set of empirical analysis along with five statistical approaches to examine learning level order will be described in this part of the dissertation. Next, chapter 4 reports the results for each study. For logical sequence, the findings will be organized into sections that show evidence to address each theoretical claim that learning progressions hypothesized. Finally, chapter 5 concludes the dissertation by discussing the results across studies and limitations of the conducted investigations. The final chapter will end the research report by outlining some future directions and summarizing a few take-away messages that were informed from the studies.

# CHAPTER II

# LITERATURE REVIEW

In this chapter, the existing literature that relates to the topic of using statistical models to examine learning progression theories will be reviewed. The chapter starts with reviewing a concept of “learning hierarchy” proposed by Gagne (1962) which can be thought of as a predecessor of learning progressions (Lobato & Walters, 2017). This concept carries some features that are similar to the newer concept of learning progressions. It is also noted that learning hierarchies had been hotly debated in the 1970s and 1980s among scholars in educational psychology, curriculum and instruction. Then, several key definitions for the concept of learning progressions will be summarized. After that, a concrete example of a theory of three learning progressions that was the baseline theory to develop an assessment system to collect the empirical data that were analyzed in this study will be presented. In the next step, the issue of validating learning progression theories will be discussed and several psychometric models that have been used to evaluate these theories will be introduced. Finally, the chapter will be concluded by a summary of the literature reviewed in this study.

## 2.1. Overview of Learning Progression Theories

To set the stage for the rest of the dissertation, this section focuses on three tasks. First, the concept of learning hierarchy and a few definitions for learning progressions will be reviewed. The former concept of learning hierarchy connects well with more recent works on learning progressions to prior investigations of learning theories from 1960s and 1970s. Second, the theory of learning progressions underlying the empirical data analyzed in this dissertation will be described in detail.  Last, two aspects of validating learning progression theories will be discussed

### 2.1.1. The "Forgotten" Concept of Learning Hierarchy

In the published works that reviewed the concept of learning progressions (e.g., Daro, Mosher, & Corcoran, 2011; Heritage, 2008), it is usually reported that the concept took root from a study by Simon (1995) in mathematics education. However, when the term "learning hierarchy" was searched for in research databases as recommended by Dr. Ronald Hambleton, numerous published works from the 1960's to early 1980's that defined and investigated learning hierarchies were found. The review of the two concepts revealed that learning hierarchies and learning progressions shared a common definitional feature in that they both capitalize on the assumption that students acquire and master knowledge and skills in a hierarchical order from simplicity to sophistication. Thus, it is worthwhile to revisit the former theory and methods used by researchers to validate the hierarchies.

Historically, Gagne (1962) laid the foundation for the term *learning hierarchy* to be coined and investigated in subsequent studies. Originated by learning psychologists and instructional designers, this concept refers to the ordered transitional relationships of knowledge elements within learning tasks. Those hierarchical relationships inferred that students need to possess the simpler elements to be able to master the more complicated ones in the hierarchy with relevant instruction (Gagne, 1962; Resnick, 1973; White, 1973). This definition is similar to that used more recently by leading authors to characterize learning progressions. After discussing the hierarchy of knowledge, Gagne (1962) introduced for the first time a hierarchy with nine elements that students went through to perform well on the task of finding the sum of a series of numbers. The hierarchy started off with the five simplest elements and proceeded to the next three elements before reaching the highest one in which students can figure out the general formula for the sum of a numeric series. The author, then, developed test items associated with each element and used them to collect response data from seven students in ninth grade to initially validate the hierarchy. He observed that the ordering of the elements from simple to sophisticated knowledge seemed to be supported by the response data. Among the seven participants, anyone who performed well on the higher-level elements, also succeeded on answering items targeting lower-level elements.

Following the model described in Gagne (1962), many researchers had attempted to propose and validate learning hierarchies in mathematics and science (White, 1973; 1974). According to White (1973), many learning hierarchies proposed and investigated in the decade following Gagne (1962) were not fully supported by empirical data. Indeed, studies to validate these hierarchies often reported non-negligible numbers or percentages of students whose response patterns were inconsistent with the prerequisite relationships of their elements. It was reported in those studies that it was possible for many students to be proficient at superordinate skills, but not the subordinate ones. Then, White (1973) pointed out three main reasons for which one could fail to validate the prerequisite relationships among the knowledge elements. These reasons were (i) the possible measurement error of the assessment instruments, (ii) the probable delay between learning and testing that might cause random forgetting, and (iii) the fallibility of the hierarchical structure. White (1973) also reported that hierarchies that were defined by intellectual skills were more likely to be supported empirically than those that relied on verbalized knowledge. Carrying this observation into a subsequent article, White (1974) proposed and illustrated a nine-step procedure to validate a learning hierarchy to maximize its plausibility. This procedure was then adopted successfully by other researchers (e.g., Winkles, 1986). Another line of research following the introduction of learning hierarchies was to investigate how to take advantage of these theories to individualize testing using computers (e.g., Ferguson, 1969) or reduce testing time at the same time with keeping an adequate level of measurement error (e.g., Spineti & Hambleton, 1977).  For instance, Spineti and Hambleton (1977), through a simulation study, found that it was possible to use learning hierarchies and adaptive testing strategies to reduce testing time by more than 50% without scarifying the level of measurement precision of conventional assessments.

Given that more modern psychometric models were in their early stages in the 1970s, the studies reviewed above that aimed to validate hierarchical learning structures based on either observed scores under a classical test theory framework or Guttman scaling (Guttman, 1944) to shed light on the plausibility of the hierarchical relationship among learning elements (Resnick, 1973). Before moving to the next topic to discuss learning progressions, it is noted that the term "learning hierarchy" tended to fade away from the scholastic discourse of K-12 education after 1990. When articles in peer-reviewed journals indexed in Educational Resources Information Center (ERIC) from 1990 to early 2018 were searched using the term in the title and "education" in any part of the publications, only 31 results were found. Moreover, most of these works were on e-learning and professional education. In short, in this section, the initiation, development and diminution of the concept of learning hierarchies were summarized.

### 2.1.2. The Concept of Learning Progressions

#### 2.1.2.1. Learning Progression.

While learning is a concept that has been around for a long time (Houwer, Barnes-Holmes, & Moors, 2013), learning progression is a much more recent idea (Lobato & Walters, 2017). Before going into the details of a few definitions for the concept, it is noted that we usually use the term “learning trajectory” in place of “learning progression” in the field of mathematics education (Confrey et al., 2017). One of the first definitions of this concept dates back to about two decades ago. In a study by Simon (1995) on constructivist education in mathematics, the author coined the term “hypothetical learning trajectory” to refer to “teachers’ prediction as to the path by which learning might proceed” (p. 135). This work is considered to be the first one that introduced the concept of learning trajectory/progression in mathematics education (Daro et al., 2011). A year later, from a measurement perspective, Masters and Foster (1996) defined learning progressions as vertical learning developments that describe knowledge and skills in a sequential order of cognition that a typical learner would go through. In the heart of this definition is the concept of learning that happens in a sequence from simple knowledge and skills to the next level of more complicated understanding and ability. A few years later, Wilson and Bertenthal (2005) proposed another definition of the concept in the case of science learning in K-12 education.  The authors defined the term as the description of “ways of thinking” about a concept that increase in the order of successive sophistication, and learners progress along the order while they learn the concept. This definition emphasizes the move of the learner from novice to expert understanding of an idea or concept.  More recently, many authors have tried to make the definition clearer and more detailed. For example, Heritage (2008) characterized a learning progression as the description of knowledge and skills that a typical student must learn in an order that helps her/him achieve more sophisticated understanding and skill sets.

Lastly, Corcoran, Mosher, and Rogat (2009) offered a definition for the concept from the perspective of empiricism. They defined learning progression in science education as an “empirically grounded and testable hypothesis” that explains how the understanding and skills of students related to a certain content knowledge develop and reach a higher cognitive level through learning activities with suitable instruction.  This definition carried some important aspects. First, it emphasized the empirical nature of the learning descriptors for each learning level. In this sense, the descriptors should be made based on empirical evidence of student learning and can be tested by observed data and appropriate techniques. Second, the definition mentioned the role of instruction in the learning development of the progression. Without appropriate instruction, student learning might not progress as the theory would predict. For example, if instruction was not aimed at helping students correct their misconception of fractions and decimals, students may keep making mistakes in adding fractions or converting fractions into decimals. More seriously, they may carry that misconception with them for a long time (Erlwanger, 1973).

After about two decades of development, there are a good number of studies of various learning progressions. A quick search by the term “learning progression” in ERIC database in the Fall of 2017 yielded 54 documents with the term in the titles. When the search is extended into “All Text” the number went up to 144 documents. When both terms “learning progression” and “learning trajectory” were used, the numbers rose up to 237 and 326, respectively. In terms of subject areas, learning progression theories have been developed for K-12 mathematics (e.g., Arieli-Attali, Wylie & Bauer, 2012; Briggs, Diaz-Bilello, Peck, Alzen, Chattergoon, & Johnson, 2015; Confrey et al., 2017; Shin, Wilson, & Choi, 2017), K-12 science (e.g., Chen, 2012; Furtak, Morrison, & Kroog, 2014; Wilson, 2009), and for verbal comprehension (e.g., Bailey & Heritage, 2008; Greaney & Tunmer, 2010). All those references contain detailed examples of learning progressions.

In terms of the number of levels within learning progressions, a closer look at the examples of learning progressions revealed that they often have a few to more than 10 learning levels. For example, Shin et al., (2017) showed an example of a learning progression in middle-school curriculum of statistics and modeling that can have two levels: being proficient and being non-proficient. On the other extreme, Briggs et al. (2015) presented an example of a learning progression for place value with up to 15 levels spanning from early pre-K to the end of grade 5. Given the incremental nature of learning progressions that can be described by levels, a formative assessment system that collects evidence of learning multiple times and provides the information for teachers and students during a course of instruction seems to be an appropriate instrument to assess learning progression of students.

#### 2.1.2.2. Learning Progression Theory.

A few learning progressions can form an educational construct and the relationship among the progression can be theorized. Under this context, a theory of learning progressions consists of (i) descriptions of each progression, and (ii) postulated relationship among them. For example, educational constructs such as mathematics proficiency in K-12 education can be viewed as multiple related learning progressions (Confrey et al., 2014), thus can be considered as theories of learning progressions. In this case, the link among learning levels across progressions within a construct can also be theorized. For instance, for a construct of two linked learning progressions of three levels each (e.g., below proficient, proficient, and advanced), it might be very likely that a student that is below proficient for the first progression also tends to be in the lowest learning level of the second learning progression. The possible occurrence of learning levels across progressions within a construct is referred to as level links or permutations or combinations of levels (Arieli-Attali et al., 2012; Shin et al., 2017; Pham, Monroe, & Wells, 2016). Those terms will be used interchangeably in this dissertation. In the next section, A specific example of a learning progression theory of middle-school mathematics will be described next.

#### 2.1.2.3. Learning Hierarchies and Progressions.

After visiting the concept of learning hierarchies and some definitions of learning progression, one could see that the definitions share a few common features. They all are likely to approach the concept from multiple perspectives. From a behavioral view, they described learning as observable phenomenon of incremental sophistication that students build the next levels of understanding on top of the previous ones. Through the definitions, we can also see that the authors emphasized the empirical aspect of how to come up with and validate those theories. This can be best seen through the last definition of Corcoran et al. (2009) since it emphasizes the importance of the empirical grounds, testable nature, and the role of instruction of the learning development. Equally important is the constructivist root of the concept of learning hierarchies by Gagne (1962), and learning progressions/trajectories from the work of Simon (1995). We encountered this constructivist facet again in the definition of the latter in Corcoran et al. (2009), and of the former in Gagne (1962) in which they both mentioned the role of instruction in how students would proceed along the knowledge and skill ladder encapsulated by the learning progressions. Suitable instruction in these definitions might be referred to what Simon (1995) described as how teachers framed their lesson plan based on their understanding of how a typical student learned the content area at hand and implemented that plan on a constructivist manner. Relevant instruction was also mentioned as a significant component of learning hierarchy theories (e.g., Gagne, 1962; Resnick, 1973; White, 1973; 1974; Winkles, 1986). In the next section, a concrete example of a learning progression theory will be shown. For examples of learning hierarchies, readers are referred to the references mentioned previously in this chapter.

### 2.1.3. An Example of a Learning Progression Theory

In this section, a theory of learning progressions for middle-school algebra proposed in Attali-Arieli et al. (2012) will be introduced. Originally, the theory contained three related learning progressions of middle-school mathematics: Equality and Variable (EV), Functions and Linear Functions (LF), and Proportional Reasoning (PR). The EV progression is integrated from two separate but related concepts that reflect students' conceptual and procedural understanding of equality and the nature of algebraic variables. LF addresses students' cognitive development of the functional relationship that starts from numeric and spatial understanding and progresses toward symbolic understanding of variables and functions. PR describes cognitive progressions that students often go through to understand the multiplicative relationship between two or more quantities. Each of the three learning progressions has five levels that describe a pattern of understanding students may pass through on their way to more sophisticated use and sense of the mathematical concepts involved. The transition from one level to the next can represent a conceptual change in understanding or a deeper understanding of an existing concept. Our previous analyses of the data supported the theory for the last two progressions (LF and PR) and failed to back up the first one (EV) (Pham et al., 2016). For brevity sake, the following paragraphs describe the theory for the LF and PR progressions. Arielli-Attali et al. (2012) contains in-depth descriptions of the theory for all the three progressions.

LF was proposed under the idea that students build their knowledge of functions from simple to more sophisticated representational understanding (Arieli-Attali et al., 2012). It starts with simple numeric and spatial representations of functional relationship and changes to more complicated graphical and symbolic representations at higher levels. For example, lower-grade middle-school students can often recognize and complement patterns of numbers such as 2, 5, 8, 11, …, and work with uncomplicated pie or bar-charts. At higher levels of the progression, more typical of later middle school, students can navigate through content knowledge that integrates numeric and visual representation such as specifying the function value associated with a variable value given the function's graph (i.e., graphical representation). Another aspect of this progression is that students gradually understand the dependent relationship and change relation between two variables (i.e., input and output). The three representational understandings (i.e., numeric, spatial, and symbolic), and the conceptual understanding of change interact to define five learning levels for this progression. In the first level, students possess the three representational understandings which are still disconnected in this stage, and they don't recognize the mutual change in this level yet. Students in the second level start to develop the concept of mutual change and integrate numeric understanding and how a pair of numbers can be represented in a two-dimensional coordinate plane. In the third level, the concept of linearity and constant change start to emerge and the first two representations are strengthened and connected with the most advanced representation of functions (i.e., symbolic). For the next level, students' understanding of the three representations of functions and their connection is crystallized. Indeed, level-4 students master the concept of constant change and they can compare the changes of different linear functions. In the most advanced level (i.e., level-5), students have the insight that how functions change might depend on the value of its variable. And, in symbolic form, students in level-5 can see that the slope of functions can vary across the range of the variable.

The PR progression was proposed by capitalizing on two major lines of research (Arieli-Attali et al., 2012). The first one was the three-stage development of the concept of proportional reasoning that took root from the work of Piaget and Inhelder (1975). These stages are (i) qualitative-intuitive, (ii) quantitative additive, and (iii) multiplicative structure (Arieli-Attali et al., 2012). The second line was from studies that reported and supported the additive misconception of students in which they use the differences between nominations and denominations instead of using multiplicative factors to compare ratios. Combining the existing theories, Arieli-Attali et al. (2012) proposed five learning levels of PR. The lowest level is associated with the qualitative-intuitive stage in which young students can make qualitative statements that compare two portions of an object. In the second level, quantitative understanding starts to emerge in students in the sense they begin to recognize the dependency of a ratio value with its components. However, the dependency understanding in this stage is still immature and students usually focus on only one part of the ratio. At level-3, a student can recognize the multiplicative nature of ratios and starts to recognize that a ratio is an independent object whose value depends on two quantities. Given this multiplicative understanding, students can map or transform one ratio to the other. However, they may still have a partial understanding that results in use additive strategy (that does not preserve the ratio) instead of the multiplicative one (that does preserve the ratio). In level-4, students can apply the multiplicative strategy correctly to transform the numerator and denominator to preserve the ratio and to solve rational problems. This means that students in level-4 grasp the functional relationship between a ratio and the two quantities that define it. In level-4, given two quantities, the students can build a ratio and keep this ratio the same by multiplying both its denominator and numerator with the same scalar. When a student can handle ratios that involve more than two quantities correctly, she/he is in level-5, the highest level of this progression.

From the definitions of LF and PR, one could see that learning levels across progressions are not independent of each other (Areli-Attali et al, 2012). For example, students with more advanced symbolic understanding of functions are more likely to be able to transform a fraction to preserve its value, thus be in level-4 of PR. Based on the descriptions of learning levels, the authors of the theory proposed a network model for the relation of the levels between the progressions. For the same reason of brevity mentioned above, only the theorized links between learning levels of LF and PR will be presented in the following. Given that each progression has five levels, there are 25 possible combinations of levels across the two progressions. However, according to the authors, it is extremely unlikely to have students in certain combinations due to the nature of pre-requisite knowledge required by the levels (Areli-Attali et al., 2012). For example, a student in level-3 of LF who can understand and work with linear functions should grasp the basics of a multiplicative relationship, thus be at least level-3 of PR. Among the 25 combinations, Arieli-Attali et al. (2012) predicted 10 possible links for LF and PR levels. Table 2.1.2 by the end of this chapter displays the postulated combinations.

In comparison to the general definitions of learning progressions reviewed earlier, it could be seen that the definition of LF and PR by Attali-Arieli et al. (2012) seems to be in line with the most recent one by Corcoran et al. (2009). Certainly, the theory of LF and PR was research-based and contains evaluable hypotheses of how students' knowledge and skills of functions, linear functions and proportional reasoning take root, accumulate and crystallize over time as they learn these concepts. In terms of connecting the theory with instructional practice, the authors of the theory have also been investigating how to assist teachers to use information informed from assessment results using items developed to measure the progression to support student learning (Wylie, Arreli-Attali, & Bauer, 2014). In short, the theory of LF and PR described previously is well-defined and ready to be evaluated. Almost all of its claims were supported by our prior studies (Pham et al., 2016; Pham et al., 2017). In this study, more analyses will be conducted to collect more validity evidence to investigate the theory from one more angle. If findings of this study confirm previous conclusions, the theory of LF and PR and items developed to validate the theory are recommended to be used to build formative assessments to support student learning and instructional practices in the area of learning and teaching those concepts. It is also noted that the empirical basis for the theory of LF and PR is generalized from cross-sectional data by observing how students learn these concepts at one point in time. One possible next phase of this study is to collect longitudinal data using the existing assessment tasks to keep track of learning trajectories of each student over time. This point will be further elaborated in the future directions following this work.

### 2.1.4. Validating Learning Progression Theories

As described in the previous sections, a typical learning progression theory that involves several related progressions often possesses two major claims. The first claim pertains to the **ordering** of learning levels within each progression. This claim anticipates that learning levels are ordered increasingly in cognitive complexity. It signifies that students in lower levels show mastery of simpler knowledge, understanding and skills of the content area defined by the progression whereas students in higher learning levels can have deeper understanding and more advanced skills than their peers in lower levels. It is also possible that the latter are more likely to suffer from some misconceptions of the concept entailed in the progression. Conversely, the former is much less likely to suffer from such misconceptions (Attali-Arieli et al., 2012). It is noted that the first claim is a natural deduction from how learning progressions have been defined. This claim has also been the main focus of most published studies that dealt with evaluating learning progressions (e.g., Kizil, 2015; Neumann, Viering, Boone, & Fischer, 2013; Steedle & Shavelson, 2009). The second claim of theories of multiple related progressions is the theoretical prediction of **the co-occurrence of levels, or level links across progressions**. This claim was described and discussed at length in Wilson (2012). Based on empirical evidence of student learning, this claim usually conjectures plausible combinations of levels. It can also be stated as one level of a progression is a prerequisite to another level of a related one. For example, of the 125 possible combinations of levels across three progressions proposed by Attali-Arieli et al. (2012), only 17 were postulated to likely occur. In this theory, students in the lowest levels (level-1 or 2) of the first two progressions (i.e., EV and LF) were not expected to master the knowledge and skills described in level-5 of PR.

Given the two claims of a typical learning progression theory, validation of such a theory requires gathering evidence to argue for or against each of the theorized assertions. If these claims are supported, we can rely on student performance on assessments based on the theories to infer practical interpretations of student learning. The validity argument for interpretive purpose of test scores encompasses multiple perspectives (Kane, 1992). In the case of building formative assessments based on learning progression theories, these aspects include, but are not limited to (i) the validity of the learning theory that defines the construct that the assessment is developed to measure, (ii) the trustworthiness of the evidence that supports our inference in the learning status of students in the progression, and (iii) the quality and usefulness of inferences regarding next learning activities that instructors can make from the assessment results to provide feedback, and set up appropriate instructional sequences. These facets can be interpreted in the language of the most recent Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National

Council for Measurement in Education, 2014). The first aspect is to collect validity evidence to examine the internal structure of the items measuring different learning levels. The second one is to evaluate the soundness of the interpretation of student knowledge and skills inferred from assessment results. And the last one is to investigate the applicability of the theory of action informed by the learning progression theory to improve student learning. The following paragraphs discuss the first aspect of the validation in details since it is the focus of this dissertation. The other two aspects will be discussed as follow-up directions for future studies in Chapter 5 of this dissertation.

To assess the first claim of level order, various sources of evidence can be informative. For example, content experts can provide feedback on the knowledge and skills that the items are written to measure each learning level (Wylie et al., 2014). Psychometric models can be also used for this validation purpose. Indeed, item parameters such as the classical or IRT-based difficulty parameter can offer some evidence regarding the complexity of the items measuring different levels (Neumann et al., 2013). This first set of evidence relates to validity evidence based on internal structure of the assessment developed to measure learning progressions. As for the second claim of co-occurrence of levels, input from content experts and curriculum studies can be a reasonable source of validity evidence to examine the possible cross level-links. This type of validity evidence associates with the content of the test, and to a lesser extent, the response processes of students when they are working on items from different progressions in the same assessment (American Educational Research Association, American Psychological Association, & National Council for Measurement in Education, 2014). Again, if some psychometric models are in use, probability-based frameworks such as those described in Pham et al. (2016) and Shin et al. (2017) might be adopted to support or negate the possibility of each combination. The following section each of the models will be review and summarized to set the stage for the next chapter on methods and statistical analyses.

## 2.2. Psychometric Models to Evaluate Learning Progression Theories

In this section, the MIRT and CDM frameworks will be reviewed briefly. The general formulation for each model will be introduced first. Then, specific versions of the models to be applied in the context of this study and their applications to evaluate learning progression theories will be detailed.

### 2.2.1. Item Response Theory

#### 2.2.1.1. IRT Models.

Under an IRT framework, the probability of an examinee to answer correctly a binary item is often a function of the examinee's proficiency and some item parameters. This function is often referred to as an item response function. A popular model of this framework is the unidimensional two-parameter model. As the name of this model suggested, each item has two parameters: the discrimination (*a-parameter)*, and the difficulty (*b-parameter*), and examinees' proficiency is characterized by one unidimensional variable (*theta*). The mathematical form of the model is as in the following:

$$P(X = 1|\theta_i) = \frac{1}{1+\exp\left(-a_j(\theta_i - b_j)\right)}, \qquad (1)$$

where $\theta_i$ is the proficiency parameter for examinee *i*, $a_j$ and $b_j$ are the discrimination and difficulty parameters of item *j*, respectively. When the discrimination parameter is fixed at a certain value for all items, the two-parameter model then becomes the one-parameter model. When the *a-parameter* is set at 1 for all items, one obtains the Rasch model. Another item feature, namely the pseudo-guessing parameter, reflects the non-zero probability that a student with extremely minimal proficiency could still answer the item correctly, can be added to equation (1) to make up the three-parameter model. To accommodate tests with polytomous items, equation (1) and its generalized version for the three-parameter model can be extended naturally to be the link functions for each response category. Since the focus of this dissertation is on formative assessments of dichotomous or dichotomized items, the two-parameter model for these items and its multidimensional model are discussed in detail. More information on other models can be found in Hambleton, Swaminathan, and Rogers (1989), Lord and Novick (1980), and Reckase (2009).

When proficiency is theorized to be multidimensional, equation (1) can be extended in several ways to parameterize the probability of endorsing an item measuring a multidimensional construct. By such expansions, one obtains multidimensional IRT (MIRT) (Reckase, 2009). Under MIRT framework, there are multiple ways to set up item response functions. These functions can reflect the compensatory nature of the latent variables in which students' higher proficiency in one dimension can compensate for their low one in the other dimension. On the other end, MIRT models can be non-compensatory in the sense that students need to have high proficiency in all the dimensions to possess higher probability to answer items correctly. Items in MIRT models can be written to measure and then loaded on only one or multiple dimensions. In this study, the MIRT-SS model in which there are two dimensions representing two progressions and each item is written to measure only one dimension (i.e., simple structure) is considered. In this model, each dimension represents one unidimensional construct defined by each learning progression. The mathematical form for the conditional probability of examinee *i* of proficiency $\theta_{li}$ for learning progression *l, (l=1 or 2)* to answer correctly item *j* measuring this progression is:

$$P\left(X_j = 1 \middle| \theta_{li}\right) = \frac{1}{1 + \exp\left(-a_j\left(\theta_{li} - b_j\right)\right)} \tag{2}$$

The notations for item parameters in this model are the same as in the case of the unidimensional two-parameter model described in equation (1) previously. It is also noted that the probability of examinee *i* to endorse item *j* depends only on her/his proficiency parameter of the progression that this item was written to measure. Thus, the MIRT-SS is a non-compensatory model. In the language of structural equation modeling (SEM), the model can be described by the path diagram in Figure 2.2.1 by the end of this chapter. In this figure, $n_1$ items from item 1.1 through 1.$n_1$ measure the first learning progression. Similarly, $n_2$ items from item 2.1 to 2.$n_2$ assess the second progression.

In the MIRT-SS model, items are assumed to differentiate students into one of two adjacent levels within each progression. For example, if the progressions were described to have three levels, namely level-1, level-2, and level-3 ordered from low to high, there will be two item groups, level 1-2, and level 2-3 to be written to help identify student learning levels. Those who did not perform well on the set of level 1-2 will likely be in level-1. On the other hand, those who show good work on those items tend to be in level-2 or 3 depending on their performance on the level 2-3 items. In this MIRT-SS, the dimensions representing two progressions are theorized to be correlated and each item is associated with only one dimension.

**2.2.1.2. Applications of IRT models in assessing learning progressions.**

Up to this moment, there were at least a dozen studies that have adopted various IRT models or their multidimensional versions to analyze learning progression data (e.g., Black, Wilson, & Yao, 2011; Chen, 2012; Chen et al., 2017; Kizil, 2015; Neumann et al., 2013; Paik, Song, Kim, & Ha, 2017; Pham et al., 2016; Shin et al., 2017).  To assess the first claim of the theories, the difficulty parameter estimates for dichotomous items and category thresholds for polytomous items were used to evaluate the ordering of learning levels. The claim of level ordering is plausible if items measuring lower levels tend to have lower difficulty estimates (Neumann et al., 2013). Pham et al. (2016) observed this pattern for two out of three progressions that the study investigated using a set of MIRT models. If polytomously scored items are used, and each score point is associated with a learning level, the threshold parameters of lower to higher categories are expected to increase from low to high. A study by Chen (2012) that investigated the usefulness of several types of polytomous items in locating student levels indicated that the thresholds of items from one type were ordered as one would expect. That study did reveal that the thresholds of a good number of their constructed response items were ordered and seemed to be useful in classify students into learning levels as one would expect. However, these findings did not hold true for most of their ordered multiple-choice and multiple true false items. Later on, Kizil (2015) reached similar conclusions for the usefulness of the ordered multiple-choice items of learning progression assessments in their study.

When the Rasch model was adopted to evaluate the ordering of learning levels using item locations, *Wright maps* were implemented to provide a visualization of item difficulty estimates and student proficiency on the same scale (e.g., Black et al., 2013). In measurement terminology, Wright maps can be considered as empirical representations of construct maps which elaborate the construct that assessments are developed to measure (Black et al., 2013). In the context of intertwined learning progressions, a construct map can be drawn for each progression, thus a Wright map can be created to visualize item and examinee locations within each progression. According to Black et al. (2013), Wright maps have some advantages in the context of formative assessments using learning progressions. First, they are a useful tool to present the assessment information in a way that is easy for teachers and students to interpret. In fact, relying on the locations of items and students in the maps, teachers can come up with adequate feedback regarding the current status of student learning and what s/he can work on in the next step to improve her/his knowledge and skills. Similarly, students can use the maps to provide feedback to their peers. Second, for a learning progression theory that involves multiple progressions, one construct map can be described to represent each progression. In this case, Wright maps for all the construct maps can be plotted side by side in a graph as illustrated in Black et al. (2012). Such graphical representation is believed to help teachers and students make better sense of student learning in reference to the theoretical constructs underlying the learning progression theory. Third, if data are collected at multiple time points, Wright maps drawn from the data enables us to describe student learning growth over time. For example, a Wright map created at the beginning of a semester indicates that a student is at learning level-2 of a progression. S/he shows a growth of two learning levels on the same progression by the end of the semester if her/his location in the Wright map drawn at the later time signifies that s/he has moved up to level-4. If longitudinal data of multiple construct maps are available, a graph of multiple panels in which each panel is a single Wright map can be created at each time point. Then, the series of graphs over all time points carries information on student learning growth in the criterion-referenced sense since the growth reflects student learning advancement regarding the domain of content described by the construct maps. In what follows, some studies will be discussed that used IRT to investigate the second claim of level links of learning progression theories.

To address the second claim, IRT models and its multidimensional versions were also likely to be useful in assessing the co-occurrence of levels of learning progressions theories. In effect, Pham et al. (2016) fit a two-tier MIRT model (Cai, 2010) to estimate the correlation of the latent variables underlying two progressions. In this study, each item was written to measure only one progression and some of them shared common stimuli, thus were in testlets. Given the testlet structure, a two-tier MIRT model that specifies latent constructs defined by testlets of items as specific dimensions, and latent variables representing the progressions as primary dimensions was in use. The model allowed the authors to estimate the correlation between the primary dimensions underlying the two progressions. Then, they used the correlation along with student proficiency estimates as well as cut scores to examine the plausibility of 10 level combinations of LF and PR. They computed model-implied probabilities and observed proportions for each combination and these statistics enabled them to support eight of the combinations. Another example of using MIRT to evaluate the second claim is the study by Shin et al (2017). In this work, the authors introduced a new parameterization by adding discontinuity parameters into the popular two-dimensional Rasch model to create change-point structured construct model (SCM-C). The purpose of the new parameter was to depict the hypothesized links between learning levels of complicated learning progression constructs. The study illustrated that the model was recoverable and obtained improved model-data fit for their empirical data. It was concluded that the change-point model was useful in supporting or disapproving the hypothesized level links using data of complicated learning progressions. It is noted that most of the dozen studies using IRT to evaluate learning progressions reviewed in this study focused on examining the first claim of level ordering. Evaluating the second claim of level links across progressions seemed to be more challenging and require more investigations. This challenge to validate level links was recognized and discussed at length in Wilson (2012). One disadvantage of IRT models in assessing learning progressions might be that they require cut scores to classify students into the levels, and thus to validate level links. In this sense, CDMs might overcome this drawback by using a probabilistic framework to place examinees into latent classes defined by learning levels.

### 2.2.2. Cognitive Diagnosis Models

As defined earlier, learning progressions are theories that describe students' knowledge and skills of a certain content area in an increasing order from simpler to more sophisticated. The increasing sophistication of learning progressions are usually phrased in descriptive learning levels. This definition is supported by some prior studies on learning that suggested it might be reasonable to characterize student learning as discrete classes (Pellegrino, Chudowsky, & Glaser, 2001). Under this view, it is recommended to fit CDMs to the learning progression data to inspect how well the items classify students into the postulated levels. Those models enable us to estimate the probability that a student is in a latent class given his/her performance on a set of items. CDM framework is a rich family of many specific models. Those models share the common goal of CDMs that classifies students into discrete latent classes. However, they vary by the number of parameters and the ways restrictions are set up among the parameters using Q-matrices (Tatsuoka, 1985). In this framework, a Q-matrix for an assessment form specifies the interaction between items and psychological attributes measured by the assessment. In this review, focus is on discussing the DINA model and its higher-order version due to their popularity and applicability to the context of evaluating learning progression theories.

To set up the mathematical form of a general CDM, Q-matrices that depict the interaction between items and attributes is needed. For an assessment of $I$ items measuring $A$ attributes, Q-matrices are of size $I \times A$ (Tatsuoka, 1985). The entry of row *i*, and column *a* of those matrices is zero if item *i* was not purported to measure attribute *a*, and becomes one, otherwise. Relying on the Q-matrix, various CDMs can be introduced to classify examinees into $C = 2^A$ possible latent classes. Each class will be denoted by a series of *A* elements of 0 and 1. The number 0 in the $a^{th}$ position indicates that students in that class do not master the $a^{th}$ attribute. Whereas, a value of 1 in that position signifies that those students show mastery of attribute *a*. For example, a student classified in a latent class encoded by [1010] masters the first and third attributes and did not master the second and the fourth ones.

Among CDMs, DINA might be the most widely used one for its simplicity and interpretability (de la Torre & Douglas, 2004). For DINA, it is required that students must master all the attributes elicited by an item to experience higher probability of answering that item correctly. Mathematically, the probability of student *j* in latent class *c* to endorse item *i* is:

$$P_{ic}(x_{ij} = 1|c) = (1 - s_i)^{\eta_{ij}} * g_i^{1-\eta_{ij}}, \tag{3}$$

where $P_{ic}(x_{ij} = 1|c)$ is the conditional probability of students in latent class *c* to endorse the item; $s_i, g_i$ are slipping and guessing parameters of the item, respectively. And, $\eta_{ij}$ is 1 if student *j* masters all the attributes required by item *i*, and 0, otherwise. In the language of Q-matrix, $\eta_{ij} = 1$ if the student shows mastery of all attributes that have 1 in the cells of the row associated to item *i*. On the one hand, the slipping parameter reflects the probability that a student with mastery of all the attributes required for item *i*, can still answer it incorrectly. On the other hand, the guessing parameter represents the chance for those who do not master all the attributes but can still endorse it.

To estimate the latent class of a respondent *j* of response vector $\boldsymbol{x_j} = (x_{ij})$ for *I* items, one relies on the *Bayes' theorem* to compute the probability $\alpha_{jc}$ of the person *j* to be in class *c* by the following formulation:

$$\alpha_{jc} = P(c|\boldsymbol{x_j}) = \frac{P(\boldsymbol{x_j} \mid c) * P(c)}{P(\boldsymbol{x_j})}, \tag{4}$$

where:

$$P(\boldsymbol{x_j}|c) = \prod_{i=1}^{I} P_{ic}^{x_{ij}} * (1 - P_{ic})^{1-x_{ij}}, \tag{5}$$

$$P(\boldsymbol{x_j}) = P(\boldsymbol{X_j} = \boldsymbol{x_j}) = \sum_{c=1}^{C} \omega_c * P(\boldsymbol{x_j}|c), \tag{6}$$

and, $P(c) = \omega_c$ is the proportion of the students in latent class *c* (Rupp, Templin, & Henson, 2010). Using equation (4), one could estimate the probabilities for each examinee being in each latent class. Based on the estimates, one could directly classify respondents into the latent classes associated with the three learning levels. CDMs have some advantages over MIRT in terms of the reliability for the classification and they require fewer items to reach an adequate level of reliability (Templin & Bradshaw, 2013).

In many applications, CDMs were fit to response data of items measuring some attributes within a broadly-defined construct that can be considered as continuous such as mathematic skills or reading proficiency (e.g., Mislevy, 1996; Tatsuoka, 1995). In these cases, it is reasonable to assume there is a continuous latent variable that underlies the attributes. De la Torre and Douglas (2004) proposed higher-order CDM framework to accommodate the assumption of a continuous latent variable that dictates the joint distribution of latent attributes. In this study, the continuous latent variable for each progression will be unidimensional to be in line with the MIRT-SS chosen to analyze the data. In general, a multidimensional continuous variable can be specified. In this higher-order model, a continuous latent variable $\theta$ is introduced to manipulate the joint distribution of latent attributes through a logistic regression model:

$$P(a|\theta_j) = \frac{1}{1+\exp\left(-(\lambda_{0a}+\lambda_{1a}*\theta_j)\right)}, \tag{7}$$

where $P(a|\theta_j)$ is the conditional probability for student *j* of continuous proficiency $\theta_j$ to master attribute $a$, $\lambda_{0a}$ and $\lambda_{1a}$ are the intercept and slope for attribute $a$, respectively. Equation (7) looks very similar to a unidimensional two-parameter IRT model in equation (1) in the sense that one can imply relative locations for different attributes from it. In fact, equation (7) can be rewritten as a two-parameter model in which the difficulty parameter associated with the attribute $a$ becomes $\frac{-\lambda_{0a}}{\lambda_{1a}}$. These location parameters can reflect a hierarchical order for the attributes. One attribute will be considered more cognitively demanding than another if its location is larger than the other. This feature of the higher-order model seems very relevant to evaluate the ordering of learning levels of learning progression theories. Attributes defined by higher levels should have larger location parameters than those measuring lower levels.

Using equation (7), one can parameterize the probability of examinee $j$ of a given proficiency of $\theta_j$ to be in a latent class $\boldsymbol{c}$ defined by the attributes. Let $\boldsymbol{c} = (a_k)_{k=1}^{A}$ be the cognitive profile. Then, one has:

$$P(\boldsymbol{c}|\theta_j) = \prod_{k=1}^{A} P(a_k|\theta_j). \qquad (8)$$

Introducing a continuous latent variable to manipulate the relationship among attributes defined by different levels of learning progressions has several advantages. First, it is a reasonable modeling framework for the attributes since they are theorized to be arranged in an increasing order of cognitive complexity. Second, the model can also offer a statistical framework to carry out an idea of creating two score scales for assessments of learning progressions (Briggs & Peck, 2015). The continuous proficiency estimate can serve as the overall score of students for the broadly-defined construct that encompasses all the progressions. On a more granular level of information, the cognitive profile of each student can be used as the growth scale in reference to how the learning progressions define student learning.

### 2.2.3. Applications of CDMs in evaluating learning progressions

As mentioned earlier, some recent studies used several CDMs to analyze learning progression data (e.g., Chen et al., 2017; Kizil, 2015; Pham et al., 2017). To address the claim about the ordering of learning levels, Chen et al. (2017) adopted Rule Space Model (RSM) (Tatsuoka, 1983), a member of the CDM family to evaluate and revise a learning progression of thermo-chemistry in high school science in China. These authors also fit IRT models to their data and used evidence from both IRT and RSM to test and revise their theory. Their study revealed that RSM can provide more detailed information about the possible learning paths within the progression than what the IRT models can offer. An interesting finding of the study was that they can use both IRT and RSM results to propose a revised version of the theory. In the revised theory, they can specify the cut scores to place students into learning levels on the continuous IRT scale and map cognitive profiles defined by the attributes into each level. Moreover, RSM enabled the authors to identify possible learning paths from one profile in a lower level to another profile in a higher one. In short, the study of Chen et al. (2017) opened a scenario in which one can use learning progression data to build two scales: one continuous and one discrete as suggested by Briggs and Peck (2015).

Another application of CDMs for assessing the ordering of levels in learning progression can be found in Kizil (2015). In this dissertation study, the author fit both IRT and CDMs to learning progression data collected from ordered multiple-choice items (OMC) that were developed to possess response options to reflect learning levels. For example, to measure a learning progression of three levels, the OMC will have three response options recoding knowledge and skills of level-1 through -3, respectively (Briggs, Alonzo, Schwab, & Wilson, 2006). Under the IRT framework, the partial credit model (IRT-PCM) was used (Masters, 1982).

From the CDM family, the author chose attribute hierarchy (Gierl, Leighton, & Hunka, 2006) and generalized diagnostic models (GDM) (von Davier, 2005). Among the three selected models, the last one tended to provide the most adequate information of model fit. This was not the case for the first two models which the author suggested that they might not be useful to evaluate the theory. On the positive side, the use of the two models under the CDM framework allowed the author to classify students into learning levels. Nonetheless, Kizil (2015) concluded that none of the models seemed to be dominantly useful in analyzing the learning progression data. The CDMs were likely to outperform the IRT one in terms of level classification and model-data fit. However, they provided a good amount of statistical evidence that was not consistent with what the theory would inform. For instance, Kizil (2015) reported that the GDM seemed to fit the best with the empirical data the author used in the study. However, the model classified more than half of the students into a cognitive profile that was inconsistent with the theoretical hierarchy of the levels of their learning progression theory. The author then discussed that this inconsistency might be due to the features and quality of ordered multiple-choice items used in their study.

To address the second claim of level-links among multiple progressions, some CDMs can be helpful. Indeed, Pham et al. (2017) fit DINA model to data of the first three learning levels of LF and PR described earlier in this Chapter. The model was fit to data of each progression to classify students into level-1, level-2 or level-3. Then, the classification of students was collected for both progressions. The percentages of students in each of the nine combinations of levels of LF and PR were tabulated and used to evaluate the plausibility of what the theorists had predicted about the co-occurrence of the levels. Using the model, all the five theorized combinations for the first three learning levels of each progression were supported. One to 33 percent of the students were observed in the combinations. The results from the DINA model also showed that all the nine pairs of levels across the two progressions were possible.

In summary, existing studies of fitting some IRT and CDMs to learning progression data conveyed mixed message about the usefulness of those models in evaluating the theories underlying the progressions. None of the reviewed studies showed evidence that supported every aspect of their respected theories. On the positive side, studies such as those of Chen (2012), Chen et al. (2017), and Pham et al. (2016) seemed to support the use of CDMs and IRT models in this context. Especially, the last two works provided evidence from fitting the models that support a good portion of the theories. On the less optimistic side, Kizil (2015) reported many challenges related to inadequate model-data fit for an IRT model, and unexpected results for CDMs when the author fit those models into his response data. A common theme among those studies was that they all suggested more research is needed to shed light on the usefulness of the models in supporting the development and evaluation of learning progression assessments. One way to investigate the comparative effectives of the models would be to conduct simulation studies in which true parameters are known and one could evaluate the models based on how well they recover the true values. To guide the simulation, I will review the mathematical relationship between IRT and CDMs in the next section.

## 2.3. The Relationship between IRT and CDMs

The utility of continuous IRT models for educational assessment data has been well-established and recognized by researchers and practitioners (Haertel, 1990; Hambleton & Jones, 1993). Meanwhile, CDMs assuming discrete latent variables might be more suitable for finer-grained analyses of learning strengths and weaknesses of students in a certain domain of content (Rupp & Templin, 2008). The two modeling frameworks are similar in some respects and yet different in others. On the one hand, both are probabilistic and confirmatory (Rupp & Templin, 2008). On the other hand, they rely on slightly different assumptions of the underlying latent variables. IRT models assume the continuity of the proficiency scale. Theoretically, proficiency estimates can be any real value from negative to positive infinity. CDMs, on the other hand, condition the probability of answering correctly an item on a finite set of latent profiles of examinees. Regardless of the differences in the assumption for the latent variables, the two models seem to be related. Indeed, Haertel (1990) stated that IRT and CDMs can be statistically equivalent in some cases. Indeed, parameter estimates for a two-parameter normal ogive IRT model would be derived from the ones of a two-latent class model if marginal maximum likelihood estimation with two quadrature points is used (Haertel, 1990). This relationship could also be seen for latent class models of more than two classes and multidimensional continuous IRT models. In item estimation, it was noted that using only a few quadrature points, which could closely approximate the integral form of the normal ogive model (Bock & Aikin, 1981), might be good enough to estimate the model parameters (Haertel, 1990). In addition, logistic and normal ogive models can be made almost identical by a simple scaling adjustment (Lord & Novick, 1968). Those sources of evidence support that IRT and CDMs might be in a close relationship under certain settings. Haertel (1990) confirmed this view by showing empirical evidence that both the normal ogive and the latent class models fit equally well to a set of empirical data. Equally important, in a comprehensive review of CDMs, Rupp and Templin (2008) stated that CDMs, being probabilistic models using categorical latent variables "can be used to approximate" their continuous IRT counterparts (p. 231).

On an empirical basis, Lee, de la Torre and Park (2012) revealed the relationship among parameter estimations of some popular models under CTT, IRT and CDM frameworks by fitting them to data of a state test of Mathematics. The authors chose three-parameter logistic model (3PL) as the IRT model, and DINA for the CDM.  In the study, they estimated two CTT parameters including the percent correct *(p+)*, and the corrected point-biserial *(d)*, three parameters of the 3PL, and the guessing *(g)*, true positive parameter *(1-s)*, and DINA-based discrimination index *(δ=1-s-g)* of the DINA model. Then, they computed the correlation between these indices for the test items. Through the computation, the authors found that the percent correct of CTT, and difficulty index of 3PL were highly correlated with the guessing and true positive estimates of DINA. The absolute values of those correlation coefficients varied from .87 to .94. However, the correlation between the 3PL discrimination parameter estimates, the corrected point-biserial, and DINA-based discrimination index was as low as .35 and .25, respectively. In the conclusion of the paper, Lee et al. (2012) acknowledged that their findings were based on empirical data of a particular assessment and should be interpreted with care. They went on to suggest that simulation studies are needed to shed light on the relationship among the models in different assessment conditions.

In short, IRT models and CDMs might appear different on the surface when one narrowly focuses on the assumption of their underlying latent variables. When one examines parameterization and parameter estimation closely, and empirical investigations are taken, greater similarity becomes apparent. The close relationship between IRT and CDMs when simple estimation methods are in use might not be practical in standard routines one currently adopts to estimate model parameters. Thus, studies how the IRT and CDMs models are similar or different in certain applications are needed to guide researchers and practitioners in using those models.

## 2.4. Summary of the Psychometric Models

Given the novelty of the learning progression concept, the literature review revealed about a dozen empirical studies that adopted IRT and/or CDMs to evaluate learning progression theories. Table 2.4.1 displays the core features of those studies which include (i) general information of the learning progressions, (ii) statistical models, (iii) key findings. The review of the studies conveyed mixed messages about the theories under consideration and the models in use. None of the studies, with the exceptions of Black et al. (2017), and Pham et al. (2017), could support all the claims of the theories. In one study, Chen (2012) found positive evidence for one item type but not the other two.  Kizil (2015) suggested that one model might be more useful than the other. Neumann et al. (2012) reported that the general ordering of learning levels in their study was supported. However, their results did not allow them to make conclusive statements about the exact number of learning levels for their progressions. By far, Chen et al. (2017) and Pham et al. (2017) seemed to be most positive about the effectiveness of both the IRT and CDMs that they used. These mixed findings seemed to be in line with the current perspectives of some leading scholars in the field about the challenges one has encountered in assessing learning progression theories (Confrey et al., 2015; Haertel et al., 2012, Wilson, 2012).

As noted earlier, researchers in mathematics education used the term learning trajectories in place of learning progression. Traditionally, those researchers have been using the CTT model to evaluate their learning trajectories (e.g., Confrey et al., 2017; Wylie et al., 2014). More recently, some of the leading scholars in this field started to fit IRT models to their data to evaluate their theories and build measurement scales (Confrey et al., 2017). It is seen that their interest in using CDMs in their works is also on the raise. According to Confrey et al. (2017), they are planning to fit some CDMs to their empirical data of learning trajectories as a next step for their study.

To conclude the section on psychometric models to analyze learning progression data, in the following, some drawbacks and advantages of MIRT-SS, DINA and HO-DINA over the Rasch model that is the model that allow us to create Wright maps will be discussed next. The purpose of the discussion is not to support the use of either one of the models or the other. It is believed that the decision to select models should be hinged on several factors which include, but are not limited to, (i) the purpose of the assessment, (ii) the purpose of the analysis, (iii) the nature of the data, and (iv) the availability of resources. In the context of learning progression assessments, using more than one model to analyze data to evaluate learning progression theories or build measurement scales for the assessments seemed to be reasonable (e.g., Chen et al., 2017; Pham et al., 2017).

For the Rasch model, the advantages of it and its multidimensional version over MIRT-SS, DINA and HO-DINA under the learning progression context could be seen through several perspectives. First, the Rasch model is more popular in educational assessment than the former ones. It was first introduced in 1960 by Georg Rasch (Rasch, 1960). Then, it was advocated to be adopted in educational measurement and quantitative psychology (e.g., Wilson, 2005; Wright & Douglas, 1975). Given the simplicity and popularity of the Rasch model, this model may be more familiar to a wider array of educational stakeholders. Thus, the use of Rasch model to evaluate and build scales for learning progression assessments might be more convenient and friendlier to researchers and users. Second, as explained previously, the Rasch model enables us to create Wright maps to put item location and student proficiency into the same scale visually. Black et al. (2012) explained that teachers and students can rely on the maps for learning progression constructs to make statistical inference of student learning for students at a given level of proficiency. For instance, it can be inferred from a Wright map that half of all the students of proficiency at the location of a dichotomous item will answer it correctly and the other half will choose the wrong answer. In addition, if a student whose proficiency is higher than an item location, s/he is more likely to answer that item correctly, thus they will be located in learning levels that are not lower than the one that the item was developed to measure. Third, from a practical point of view, the Rasch model, and its development to accommodate polytomous data can be estimated by many widely-used software programs.

For the advantages of the Rasch model to be realized, it is essential that the assumptions underlying the model are met and the model fits well with the data. Using the Rasch model for one construct map that represents one progression, one assumes that student learning status of this construct is continuous and unidimensional. This assumption implies that students in a higher learning levels are obviously proficient in the lower levels. However, previous investigations of learning hierarchies revealed that instructional delay or random forgetting can cause the phenomenon in which students can master superordinate elements but fail to perform well on the lower ones in the hierarchies (White, 1973). In recent years, some leading scholars also expressed their concerns about the linearity of learning progressions (e.g., Confrey et al., 2017; Kingston, Broaddus, & Lao, 2015; Lobato & Walters, 2017). If the assumption that a student who masters a higher level is automatically proficient in all the levels lower than that one is not always viable for all students, the use of the Rasch model would superimpose a linear structure of learning into a nonlinear construct. In terms of model-data fit, the Rasch model is less likely to fit empirical data as well as its 2PL or 3PL counterparts since it restricts each item to possess only one parameter (e.g., Hambleton & Jones, 1993; Sinharay & Haberman, 2014).

For MIRT-SS, DINO, and HO-DINA, the three models selected in this study have a few advantages over the Rasch model in the context of analyzing data to validate learning progression theories and build assessments based on these theories. In the first place, the baseline model of MIRT-SS is a 2PL unidimensional model which allows item discrimination parameters to take any positive value. This modeling setting is expected to result in better model-data fit than the Rasch model since it freely estimates the discrimination parameter. In one of our previous investigation of the empirical data used in this dissertation, a two-tier 2PL MIRT model (Cai, 2010) was fit to the data. It was found that most of the discrimination parameter estimates of LF and PR items ranged from .10 to 5. The mean and standard deviation for these discrimination estimates were 1.60 and .64 for LF, respectively. These statistics for PR were higher at 1.86 and .94 (Pham et al., 2016). As a result, if the discrimination parameter of the items in this study is constrained to 1, it might introduce error to the estimates of item difficulties which are essential in evaluating learning progression theories. Thus, 2PL nature of MIRT-SS warrants better model-data fit than what the multidimensional Rasch model could bring about. It should also be noted that an item-person map that is similar to the Wright map can be constructed for any unidimensional IRT model. Given that the baseline structure of MIRT-SS is an 2PL unidimensional model for each progression, it is possible to build a construct map that displays examinees' proficiency and items' difficulty on the same scale as the case of Wright map for the Rasch model.

In the second place, DINA does not assume the unidimensionality of the assessment data. By loosening this strong assumption out of the modeling framework, DINA can fit better with empirical data in which the unidimensional assumption is severely violated as what we can observed in the dimensional investigation of learning progression data by Fu, Chung and Wise

(2013) and Kizil (2015). Equally important, fitting DINA to learning progression data using appropriate Q-matrix enables us to place students into cognitive profiles defined by attributes associated with learning levels as described previously in this chapter. Each profile is coded as a set of 0 and 1 in which 0 indicates non-mastery and 1 implies mastery. The coding of the profiles supports us to make fine-grained inference of student learning status. For example, for a learning progression of three levels, two item groups to distinguish students into level-1 or higher, and level-3 or lower, respectively, can be used to define two attributes. Students classified in profile [01] are those who master higher levels (i.e., level-2 and 3), but might have forgotten some knowledge and skills required to perform well on level-1. In comparison to the Rasch model or MIRT-SS, the output of fitting DINA to learning progression data provides more diagnostic information regarding student performance on each of the specific learning levels. In other words, classification of students into learning profiles by DINA is more informative and granular than the single proficiency score by the continuous IRT models. DINA can also accommodate the nonlinearity of learning phenomenon in which students can forget prior knowledge or skills, thus master higher learning levels but perform less well on the lower ones. In the third place, HO-DINA inherits the advantage of DINA that both CDMs can classify students into learning profiles without imposing that students must not forget what they have learned previously. Additionally, through making an extra assumption of a continuous latent variable underlying the attributes measured by assessments, HO-DINA supplies us with a tool set of using attribute locations to examine the hierarchy of learning levels. More importantly, HO-DINA can offer each student with two measures: (i) the continuous proficiency as in the case of IRT models, and (ii) discrete cognitive profiles associated with the learning levels. Those two measures seem to be useful to build two scales that reflect (i) student proficiency in a broad domain of content, and

(ii) student learning growth in reference to a solid theory of learning progressions defined within the domain, respectively as suggested by Briggs and Peck (2015).

Finally, DINA and HO-DINA can be more efficient in building formative assessments based on learning progression theories. In effect, those CDMs usually require a smaller number of items to reach an adequate level of measurement reliability than what a typical IRT model would need to obtain the same amount of reliability (Templin & Bradshaw, 2013). Given the reality that formative assessments are often shorter in time and contain much fewer items, DINA and HO-DINA as well as other potential CDMs seem to be good choices to calibrate data of formative assessments based on learning theories.

In summary, each model reviewed and discussed in this chapter has advantages and drawbacks in the context of analyzing data to validate learning progression data or build assessment scales to measure student learning. To select which model should be used in each empirical analysis is not an easy decision to make. However, one can see that more recent studies to evaluate learning progressions tended to fit more than one model to the same data (e.g., Chen et al., 2017; Kizil, 2015; Pham et al., 2017). Given the increasing interest in using IRT and CDMs to investigate learning progressions/trajectories, this study is expected to shed some light on the effectiveness of those models in analyzing learning progression data. In the next chapter, detailed description of the methods that will be used to address the problem of how MIRT-SS, DINA, and HO-DINA are effective at evaluating learning progression theories will be provided.

## 2.5. Tables and Figures for Chapter 2

**Table 2.1.2. Ten Possible Combinations of LF and PR Learning Levels**

| Level | | Combinations | Level Description | |
|---|---|---|---|---|
| LF | PR | | Functions & Linear Functions | Proportional Reasoning |
| 1 | 1 | (1,1) | Separate numeric & spatial understandings | Additive-intuitive understanding |
| 1 | 2 | (1,2) | Separate numeric & spatial understandings | Start of quantitative understanding and working with single ratio |
| 2 | 2 | (2,2) | Understanding of mutual dependent change | Start of quantitative understanding and working with single ratio |
| 2 | 3 | (2,3) | Understanding of mutual dependent change | Begin to recognize multiplicative relationship |
| 3 | 3 | (3,3) | Understand and be able to work with linear functions | Begin to recognize multiplicative relationship |
| 3 | 4 | (3,4) | Understand and be able to work with linear functions | Understand correctly and work effectively with multiplicative relationship |
| 4 | 4 | (4,4) | Be able to compare constant change and linear functions | Understand correctly and work effectively with multiplicative relationship |
| 4 | 5 | (4,5) | Be able to compare constant change and linear functions | Be able to work with ratios of more than two quantities |
| 5 | 4 | (5,4) | Understand changing changes | Understand correctly and work effectively with multiplicative relationship |
| 5 | 5 | (5,5) | Understand changing changes | Be able to work with ratios of more than two quantities |

**Table 2.4.1. Summary of Existing Studies to Evaluate Learning Progressions**

| Study | Learning Progressions | Data/ Models | Key Findings |
|---|---|---|---|
| Black et al. (2012) | Science, Middle-school | MCQ, Open-Ended, Rasch | The items measuring the levels tended to be in correct order; |
| Chen (2012) | Science, Middle-school | Several item formats,<br>IRT-PCM | Thresholds of constructed response items were ordered; the model was useful. |
| Chen et al., (2017) | Chemistry, High school | IRT-Rasch, CDM-RSM | The models were helpful in validating the theory and suggesting revisions. |
| Kizil (2015) | Science, High school | Ordered Multiple-choice Items<br>IRT-PCM, CDMs | Neither of the models was found to have adequate model-data fit. |
| Neumann et al., (2012) | Science, Middle-school | MCQ, Rasch | The general ordering was supported, the number of levels was in doubt |
| Paik et al., (2017) | Science, K-12 | Open-Ended, Rasch-PCM | The progressions were supported, two dimensional Rasch model fit well with the data. |
| Pham et al. (2016) | EV, LF and PR; 5 levels;<br>Middle-school Mathematics | Dichotomous, Polytomous Items,<br>MIRT | The model was helpful to support two out of the three progressions. |
| Pham et al. (2017) | LF, PR; 3 levels | Dichotomous, Polytomous Items<br>IRT, CDM-DINA | IRT and DINA seemed to be helpful in evaluating the theory. They provided both convergent and divergent evidence. |
| Shin et al., (2017) | Statistics & Measurement;<br>Middle-school | Dichotomous Items,<br>SCM-C | The model was proved to be helpful. The theoretical claims were supported. |
| Steedle & Shavelson (2009) | Physics; Middle-high schools | Ordered Multiple-choice Items,<br>CDM | The study found it challenging to validate the theory. |

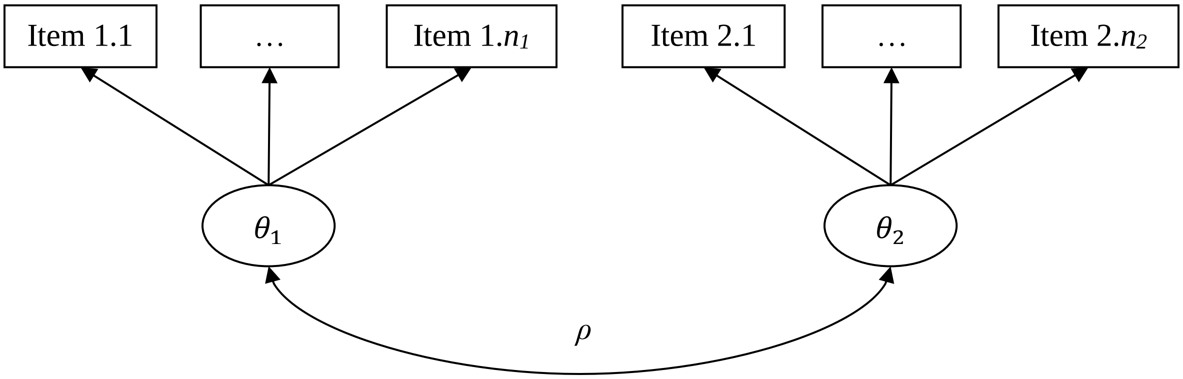


**Figure 2.2.1. Path diagram of the MIRT-SS**

# CHAPTER III

# METHOD

In the previous chapters, the topic of learning progressions was introduced, and work was reviewed that aimed to evaluate learning progression theories in the hope of building learning assessments based on the concept of learning progressions. In Chapter 1, the research purpose of examining the effectiveness of three psychometric models in analyzing learning progression data was stated. This chapter will discuss the method that was adopted to address the research problem for this dissertation. To investigate the effectiveness of the MIRT-SS, DINA, and HO-DINA models in analyzing learning progression data, two simulation studies (Studies 1 and 2) and one empirical analysis (Study 3) will be described. The purpose of the simulation studies is to evaluate the effectiveness of the three models in examining the ordering of learning levels and the theory-informed links between the levels. The data generation was designed to mimic the key features of a real assessment system constructed to collect data to evaluate a theory of learning progressions. Those features include, but are not limited to, the number of items, item parameters, and the number of examinees. To investigate the sensitivity of the models in detecting fallible theories, two scenarios, one in which the learning progression theory is true and the other in which the theory is false, are considered. They will be called true and false scenarios, respectively. In the empirical analysis, the models will be fit to real data with the purpose of determining if the models provide the same results when the data do not follow nor satisfy the assumptions of a specific model. In the following sections, each study will be described in detail.

### 3.1. Study 1: MIRT-SS as the Generating Model

In Study 1, it is assumed that learning progressions data fit perfectly with a two-parameter MIRT-SS model of two continuous latent factors that represented two progressions. Items in each progression are loaded only on the factor that corresponds to the progression. As introduced earlier, data were generated for two scenarios. In the true one, an assumption is made that the learning progression theory holds in the sense that items measuring lower learning levels were notably easier than their counterparts of higher levels. In the second scenario, the assumption of the ordering of items measuring different levels was violated. It means that difficulty parameters of items measuring different learning levels were sampled from the same distribution. Once the data were generated in each scenario, all the three models were fit to the data, and parameter estimates were analyzed to shed light on how the use of the models enabled us to detect the ordering of levels and plausibility of level links.

#### 3.1.1. Data Generation

In this study, MIRT-SS model was used to generate dichotomous data for two learning progressions (LP1 and LP2), each with three levels. learning progressions of three levels are chosen since these progressions are quite popular among the ones with more than two levels (Shin et al., 2017). Moreover, the progressions that were investigated in Study 3 also have three learning levels. In each scenario, the data were generated under three fully crossed factors: number of items that discriminate between adjacent levels (10 and 15), sample size (500 and 1,000), and proficiency correlation between LP1 and LP2 (.6 and .9). The conditions for the number of items were selected to represent typical test lengths of 40 to 60 items per test form. Although the test lengths seem to be long for formative assessments, it is noted that the purpose of this study is to investigate the effectiveness of three models in analyzing data to evaluate theories of learning progressions, not the actual applications of the assessments in a formative assessment. In this sense, 10 to 15 items per item group are typical numbers of an item bank developed to evaluate a learning progression theory (e.g., Pham et al., 2016). In addition, a power analysis conducted using G*Power (Faul, Erdfelder, Buchner, & Lang, 2009) revealed that approximately 14 items per group are needed to maintain a power of .80 to detect the item difficulty difference between two item groups of effect size of 1. Thus, two numbers of items per item groups of 10 and 15 were chosen to include the value of 14. The sample sizes of 500 and 1,000 represent small to large sample sizes reported in typical learning progression studies (e.g., Confrey et al., 2017; Pham et al., 2016; Shin et al., 2017).

In the true scenario where the learning progression theory holds, the difficulty parameter values of items in item group 1 that supports us to locate students into levels-1 or -2 for both LP1 and LP2 were sampled from a normal distribution with a mean of -.50 and standard deviation of 1.00. For items measuring higher levels (i.e., level-2 and -3), their difficulty parameter values were sampled from a normal distribution with a mean of .50 and unit standard deviation. In the false scenario, item difficulty parameter values of all the items were sampled from a standard normal distribution (i.e., the mean of the item difficulty parameters did not differ between levels).

In both scenarios, the discrimination parameter values were sampled from a log-normal distribution with the mean and standard deviation of the variable's natural logarithm were $\mu$=0, and $\sigma$ = 0.25, respectively. In this distribution, more than 99% of the *a*-parameters varied from .5 to 2. As for the student proficiency, the parameter values in both scenarios were sampled from bivariate distributions with a correlation of .6 or .9. In one extreme, the coefficient of .9 is to reflect the very high correlation of .89 between LF and PR found in our previous study (Pham et al., 2016). In the other extreme, a coefficient of .6 was chosen to represent a moderate correlation between educational constructs measured by popular tests (Frey & Detterman, 2004). A correlation of .6 also seems to be realistic in the context of learning progressions. Indeed, in an empirical study of science learning progressions, Black et al. (2012) reported a correlation of .68 between two constructs of melting and evaporation within their theory. Under this design, there were 16 conditions defined by two scenarios (i.e., true or false), two sample sizes (i.e., $N$=500, or 1, 000), two numbers of items (i.e., $I$=10 or 15 per group), and two bivariate distributions of student proficiency (i.e., $\rho$=.6 or .9). For each condition, 100 replications were performed. R was used to simulate data (R Core Team, 2017).

### 3.1.2. Parameter Estimation

The MIRT-SS, DINA, and HO-DINA models were fit to the simulated data. To calibrate data by MIRT-SS, flexMIRT (Cai, 2015) was used to estimate item parameters and student proficiency scores for each progression. These estimates provided baseline information to examine the ordering of learning levels and the plausibility of level combinations for both progressions.

For DINA and HO-DINA, the GDINA R-package was used (Ma & de la Torre, 2017) to fit the models to the data. In these calibrations, there were two attributes defined by two groups of items. The first one was the knowledge and skills defined by level-1 and -2 of the progression. The second attribute reflected the construct encapsulated by level-2 and -3. In the language of Q-matrix, items in item group 1 measuring level-1 and -2 required only attribute-1 to be mastered by students so that the students can have a higher probability of endorsing the items. Whereas, items in item group 2 targeting level-2 and 3 required students to master attribute-2 to increase the probability of answering correctly those items. The Q-matrix for one learning progression in this study was of the form of Table 3.1.1 by the end of this chapter. In this table, item 1.1 to item 1.$n_{1\text{-}2}$ were from item group level 1, whereas, item 2.1 to item 2.$n_{2\text{-}3}$ belonged to item group 2. In this simulation study, $n_{1-2}$ was equal to $n_{2-3}$, and they were the number of items per item group. They can be 10 or 15 as described earlier. The first group contains items that were written to classify examinees into level-1 or -2. Meanwhile, items in the second group measured content designated by level-2 and -3. These items aimed to distinguish students into level-2 or -3. In the language of CDM, students who do not master attribute-1 are in learning level-1, master both attributes are in level-3. And, those who master the first attribute but not the second are in the middle level.

In this setting, one has four latent classes or cognitive profiles defined by the mastery levels for the two attributes. As introduced previously, number 1 is used to indicate that a student masters an attribute, and number 0 is in use otherwise. Using these 0/1 indicators for two attributes, the four profiles can be coded as (i) [00], (ii) [01], (iii) [10], and (iv) [11]. In these notations, the first and second indexes correspond to the first and second attributes, respectively. For example, cognitive profile [10] implies that students in this class master the first attribute and don't show mastery of the second one. Among the four possible profiles, three of them excluding that of [01] reflect the three learning levels of the progression. Indeed, class [00] represents level-1 since students in this class don't show mastery for all the attributes, thus s/he should be in the lowest learning levels. Meanwhile, class [10] signifies level-2 because students in this profile master only the lower attribute that dictates level-1 and -2. The last one [11] corresponds tolevel-3due to the fact that students in this profile master all the two attributes, thus they should be proficient on all the knowledge and skills defined by all the levels. Students classified in profile of [01] master only the higher attribute but not the lower one. If the levels are ordered according to the theory, thus the attributes are in hierarchical order from low to high, it is unlikely that we will observe a significant number of students in this profile. However, if instruction of knowledge and skills of lower levels happened too far from the testing time, and that of the higher levels is more recent, students can show mastery of the higher levels but not the lower ones. Under the assumption that the theory of learning progressions holds, this profile of [01] does not represent any learning levels and would be considered inconsistent with the theory.

To estimate DINA, a prior distribution for the four cognitive profiles defined by the two attributes needed to be specified. By default, GDINA employs a uniform joint distribution as the initial distribution for the profiles and this default prior was used to estimate the slipping and guessing item parameters as well as cognitive profiles of students. In the HO-DINA, a two-parameter (2-PL) model was used to parameterize the joint distribution of the two attributes. The 2-PL model was specified, since it was the parameterization used in MIRT-SS to simulate data for this study. By default, a standard normal distribution for the continuous latent variable of HO-DINA was used for the estimation. Outputs of fitting HO-DINA to the data included (i) intercepts and slopes of the two attributes, (ii) slipping and guessing parameters of each item, (iii) cognitive profiles for each student. These outputs were used to evaluate the effectiveness of HO-DINA in analyzing learning progression data.

### 3.1.3. Data Analysis

The item and examinee parameter estimates obtained from fitting each model to simulated data were used to investigate the ordering of learning levels and proportions of students in each combination of levels across two progressions. The following sections explain the details of analyzing the data using MIRT-SS, HO-DINA and DINA.

To assess if learning levels within each progression were correctly ordered, two methods to analyze results of MIRT-SS calibrations, two methods by fitting HO-DINA to the simulated data, and one method for DINA were used. Starting with the MIRT model, two independent t-tests (one-tailed) were conducted to compare item difficulty estimates by MIRT-SS for items measuring Levels 2-3 and Levels 1-2 for two progressions. A conventional alpha level of .05 was chosen for this one t-test which was carried out for each progression. The one-tailed test was preferred over the two-tailed since the aim was rejecting the null hypothesis of incorrect order of learning levels in this study. Using conventional notations, our null hypothesis is $H_0: \mu_{\beta_1} \geq \mu_{\beta_2}$, and the alternative hypothesis becomes $H_1: \mu_{\beta_1} < \mu_{\beta_2}$, where $\beta_1$ and $\beta_2$ represent item difficulty parameters of item groups 1 and 2, respectively. As reported in the literature review of this dissertation, this method of comparing item difficulty was widely used in existing studies to evaluate learning progressions using CTT or IRT (e.g., Chen et al., 2017; Neumann et al., 2013; Pham et al., 2016; Wylie et al., 2014). To collect more information about the usefulness of this model, a test of ordered-cuts or order-test for short in which the median difficulty estimates of items from two groups of items measuring Levels 1-2 and Levels 2-3 were compared was also considered. For the t-tests, the increasing order of the learning levels of the replication was supported if both tests for the progressions yielded significant results at a conventional alpha level of .05. For the order-test, the claim was supported if the medians of difficulty estimates of the items measuring lower levels were smaller than those of the higher ones. The order-test was investigated since one doesn’t usually have many items for some learning levels in empirical studies (e.g., Chen et al., 2017; Pham et al., 2017). Under this constrain, the t-test is likely to have low power to detect the significant difference, thus the order-test tends to be more helpful in this case.

For CDMs, a framework was adopted introduced in Pham et al. (2017) to analyze the results of fitting HO-DINA and DINA to the data. For DINA and HO-DINA, student cognitive profiles for data sets whose calibrations converged normally are tabulated for each condition. As a result of fitting those models to the data, one can locate each student in one of four profiles [00], [10], [11], and [01]. The first three latent classes associate with learning level-1, -2, and -3, respectively. Whereas, the last profile doesn’t correspond to any levels. For each replication, the effectiveness of the DINA and HO-DINA models in recovering the true information used to simulate data was evaluated by two methods. The first one was the minimum test that compares the proportion of students in the inconsistent profile [01] with that of the remaining profiles consistent with the learning theory. This comparison can be done for both DINA and HO-DINA. If the proportion in the inconsistent profile was smaller than those of the three consistent levels, the minimum test yielded positive result, and it can be concluded the first claim of the theory about the ordering of the items can be supported. The second method under the CDM framework was the location test in which the locations of attribute-1 and 2 by HO-DINA were compared. If the location of attribute-1 was smaller than that of attribute-2, one can affirm that the ordering of the levels is supported. For more detail, the locations of attribute-1 and 2 for each replication was collected. Then, they were used to shed light on the plausible ordering of levels in each case. To account for sampling error, the percentages of replications whose test results were positive using each of the five methods were aggregated and reported as true positive and false positive rates for the true and false scenarios, respectively. These terms are used instead of power and type I error rates, since among the approaches, only the first is truly a hypothesis test. The remaining methods are simply binary classification tests, because the sampling distribution of the null hypothesis for these tests is not known. Given the nature of the theories in the true and the false scenarios, it is expected that the first claim is supported in the former and remains untenable in the latter.

To address the second claim of the simulated learning progression theory, only cases in the true scenario, where the ordering of levels was supported statistically or the cuts were in increasing order, were considered. The term partially-supported cases was used to indicate such eventualities. For these replications, due to the quite small number of items in each item group (i.e., 10 and 15), the median of difficulty estimates of items in item group 1 were used as the cut score to place students in level-1 or higher levels. Similarly, the median of item difficulty of items in item group 2 became the cut score for level-3 or lower levels. Then, the two cut scores were used to classify students into one of the three levels. If a student proficiency score was lower than the first cut, she/he was classified in level-1. If the score was higher than or equaled the second cut, her/his learning level was level-3. Lastly, when the score wasn't less than the first cut and was smaller than the second one, she/he was in level-2. Once the learning level for each student in each progression was identified, observed proportions of students in each combination of levels across progressions were computed. The observed proportions of students in each combination of levels informed us about how likely each combination would be for progressions using each model. The observed proportions for each partially-supported case were stored. To aggregate these statistics, averages of the proportions for each combination over all the partially-supported cases were computed for each simulation condition.

As discussed in Chapter 2, using MIRT-SS one can classify simulated students into learning level-1, level-2, or level-3 for each progression. In addition to these levels, HO-DINA and DINA sometimes will place students in the inconsistent profile [01]. In this study, MIRT-SS was used to generate response data. As a result, the known parameters for this model were adopted to located students into true classification. This classification using true item and student parameters was then treated as the baseline information to evaluate the classification accuracy by the models. To be able to use MIRT-SS parameter estimates to classify students into learning levels, it is required at the least that the cut scores to distinguish levels are in a correct order.

Given that requirement, only the true scenario was considered and replications where the cuts for each progression were increasingly ordered were taken into consideration.

To investigate the usefulness of the models in recovering the true classification into combinations of levels, classification accuracy rates for the models were averaged across correctly-ordered replications in each condition of the true scenario. In addition, cross-model classification consistency of classifying students into level combinations between pairs of models was also aggregated across these cases. The cross-model consistency rate for two models was computed as the proportion of students classified into the same combination of levels by the models. The higher the level of classification accuracy, the better the model was at recovering the true classification. In a similar vein, the higher the cross-model classification consistency, the more similar the level classifications by the models were.

Table 3.1.3 at the end of this chapter synthesizes the simulation conditions and is the key to data analysis for this study. In summation, 16 conditions defined by two scenarios (i.e., the learning progression theory holds, and doesn't hold), two sample sizes (*N*), two numbers of items (*I*) and two correlation coefficients ($\rho$) between the two dimensions ($\theta_1$, and $\theta_2$) were considered. After assessing the convergence of calibrations by MIRT-SS, HO-DINA and DINA, results from data of normal convergence were analyzed to shed light on how well the models recover the true information used to generate the data. In more detail, item difficulty under MIRT-SS framework, the percentages of students classified in the inconsistent profile of [01] for CDMs, and the attributes' locations in the HO-DINA framework were used as inputs to evaluate the first claim about the ordering of learning levels. Regarding the second claim as to level combinations, the cross-model consistency of level classifications by DINA and HO-DINA in comparison to that by MIRT-SS was assessed to elucidate how well the CDMs recover the combinations of levels of students using MIRT-SS. In summary, Study 1 assumes that MIRT-SS fits perfectly with learning progression data and the analysis based on this model correctly classifies students into learning levels. In this study, the sensitivity of the model under investigation in detecting the fallible ordering of levels can be examined using the true and false scenarios. In the next study, the generating model was switched to HO-DINA to investigate the performance of DINA and MIRT-SS on data by the higher-order model.

### 3.2. Study 2: HO-DINA as the Generating Model

Given that we never truly know which psychometric model fits perfectly with an empirical data set of learning progressions, the consideration of a generating model other than the MIRT-SS as in Study 1 is reasonable. Therefore, the HO-DINA model was used to simulate the data in Study 2. This model was chosen as the generating model for two reasons. First, it is noted that the underlying continuous latent variable in this model is likely to represent a construct that is broader than the specific content measured by items of CDM-based assessments (de la Torre & Douglas, 2004). For example, the items measuring multiple learning progressions are usually part of an item bank that assesses a broader content area defined by an educational curriculum. Second, it is observed that testing data of popular assessments measuring mathematics or reading proficiency can be considered unidimensional (e.g., Dorans & Lawrence, 1987; Robin, Bejar, Liang, & Rijmen, 2016; Zwick, 1987). Thus, the knowledge and skills defined by LP1 and LP2 can be assumed to be a part of a larger construct that is essentially unidimentional.

Similar to the previous investigation, two scenarios (i.e., true and false) were considered in Study 2. In the true scenario, locations of attribute-1 for both progressions were smaller than those of attribute-2. This setting reflects the ordering of learning levels in a true learning progression theory. To investigate the sensitivity of the models in detecting the order of levels, the magnitude of the distance between locations of the attributes was purposefully varied. In the extreme difference conditions, the distance was two logits, whereas in the moderate difference conditions, a difference of .50 logits was used. The former value was a principled choice, since a prior study by Pham et al. (2017) found that the distance between attributes observed in six empirical data sets ranged from -.05 to .73 and the median of these distances was .28. The distance of 2 logits was then selected to distinguish from the largest empirical distance of .73. For the false scenario, the locations of two attributes were set to be the same, which signifies that the learning levels were not ordered in increasing cognitive demand for knowledge and skills.

### 3.2.1. Data Generation

A Q-matrix (shown in Table 3.1.1) was used to generate and calibrate data throughout simulation conditions in this study. Identical to Study 1, two conditions of sample sizes of *N* = 500, and 1,000 and two total numbers of items (*I* = 40 and 60) were manipulated in each scenario. Being consistent with previous notations, let $\theta_1$ and $\theta_2$ represent the two continuous latent variables underlying the two learning progressions under investigation. Given a student of continuous scores ($\theta_1$, $\theta_2$) for LP1 and LP2, using equation (9) below, we can compute the probability for her/him to master an attribute $(a_k)_{k=1,2}$ of one of the progressions. To be more explicit, the following formula details the probabilistic relationship:

$$P(a_k|\theta_l) = \frac{1}{1 + \exp\left(-\left(\lambda_{0a_k} + \lambda_{1a_k} * \theta_l\right)\right)}, \tag{9}$$

where $a_{k=1,2}$ represent the two attributes, and $l=1,2$ denote the two dimensions underlying the progressions. To use HO-DINA to generate data for two progressions, several features of the models needed to be specified. They included (i) the bivariate distribution for $(\theta_1, \theta_2)$, (ii) the intercepts $(\lambda_{0a_k})_{k=1,2}$, and the slopes $(\lambda_{1a_k})_{k=1,2}$ for the attributes, (iii) and the slipping $(s_i)_{i=1 \dots I}$ and guessing parameters $(g_i)_{i=1 \dots I}$ for the items. Similar to Study 1, two bivariate standard normal distributions of correlation coefficients of .6, and .9 were chosen for the continuous latent variables ($\theta_1$, $\theta_2$). To reflect the trustworthiness of the true and false learning progression theories in two scenarios, the intercepts and slopes of attributes as well as the slipping and guessing parameters of items were manipulated in each scenario. The following paragraphs will provide the details for the parameters in each case.

In the true scenario, attribute-1 elicits lower cognitive demand than attribute-2. When the distance between attributes were extremely large, the intercept and slope of the former were all fixed at 1.7, and those of the latter were at -1.7 and 1.7, respectively. Those values were chosen so that the attributes can be put in the normal scale with D=1.7 as the scale that was in use in the original work by de la Torre and Douglas (2004). Under this setting, the location for attribute-1 and 2 are -1 and 1, respectively. To reduce the distance to .5 logits, the slope was fixed at 1.7 for both attributes and change their intercepts to .425 for the first one and -.425 for the second one. As a result, the location of attribute-1 was then -.25, and that of attribute-2 became .25. At the next step, the guessing and slipping parameters for the items needed to be specified so that those parameters mirrored the increasing cognitive demand of items targeting level-1 and 2, and level-2 and 3. Indeed, when the percentages of students mastering and non-mastering all the attributes required by an item were equal (i.e., 50%), the expected proportion correct in CTT sense of the item was reflected in the delta index $\delta=(1-s+g)/2$ (Lee et al., 2012). When the locations of attribute-1 and 2 were at -1 and 1, and with D=1.7, the percentages of students sampled from a standard normal distribution who master the first attribute is 76%, and the percentage for the second one is 24%, respectively. These percentages for the moderate difference cases are 57%

and 43% for attributes 1 and 2, respectively. Appendix 1 shows the computation for these figures. With these mastery percentages, the expected proportion correct for an item in item group 1 that requires attribute-1 is $\delta_{1e} = .76*(1\text{-}s) + .24*g$, and an item in item group 2 that requires attribute-2 is $\delta_{2e} = .24*(1\text{-}s) + .76*g$ for cases of large location distance. These formulas for the moderate difference data are $\delta_{1m} = .57*(1\text{-}s) + .43*g$, and $\delta_{2m} = .43*(1\text{-}s) + .57*g$. When *s* and *g* vary between 0 to .4, $\delta_{1e}$ ranges from .46 to .86, and $\delta_{2e}$ receives values in between .14 and .54. The intervals for $\delta_{1m}$ and $\delta_{2m}$ are [.34, .74] and [.26, .66], respectively. The computations suggest that items measuring lower levels are clearly easier than their counterparts that measure higher levels when the attributes are very distant from each other. The easiness of the items becomes less salient when the attributes are moderately distant. Nonetheless, both guessing and slipping parameters are expected to vary from 0 to .4 under CDM framework (de la Torre & Douglas, 2004). Given the reasons shown above, the decision was made to randomly select 1- *s* and *g* for items of both groups from 4-*Beta*(0.6, 1, 2, 1), 4-*Beta*(0, 0.4, 1, 2) distributions. These distributions are similar to the priors de la Torre and Douglas (2004) used to estimate item parameters of HO-DINA in their simulation study. The only difference here was the range for *g*, 1-*s*. In this study, *s* and *g* varied from 0 to 0.4, thus 1-*s* ranged from 0.6 to 1. In de la Torre and Douglas (2004), the authors used 4-*Beta*(0.4, 1, 2, 1) and 4-*Beta*(0.6, 1, 2, 1) as the prior for 1-*s*, and *g*, respectively. For practical purposes, both *s* and *g* were limited within the range of 0 to .4, since a guessing or slipping parameter higher than .4 seems to indicate poor fit (de la Torre & Douglas, 2004). This range of [0, .4] is also typical for the parameters observed in empirical data (e.g., de la Torre & Douglas, 2004; Lee et al., 2012)

In the false scenario, both locations of the attributes were set at 0 by fixing their intercepts at 0 and their slopes at 1.7. In this case, the proportion of students who master each attribute was 50%. Thus, the expected proportion correct for an item in either item group 1 or 2 is $\delta=(1\text{-}s+g)/2$, where $s$ and $g$ are slipping and guessing parameters of the item (Lee et al., 2012). Similar to the true scenario, if both $s$ and $g$ varied between 0 and .4, the delta index of items was within the range of .3 and .7. This analysis suggested to use the same distributions to sample 1-$s$ and $g$ for items in both item group 1 and 2 in this false scenario. In other words, 1-$s$ and $g$ for all the items were sampled from 4-*Beta*(0.6, 1, 2, 1), 4-*Beta*(0, 0.4, 1, 2), respectively.

Once those parameters are selected, student continuous scores ($\theta_1$, $\theta_2$) were drawn from the two bivariate standard normal distributions described earlier. In the next step, data were generated for each progression. Let $\theta^j$ be the continuous score of student $j$ for one progression. As adopted in de la Torre and Douglas (2004), a respondent's mastery profile for each attribute was drawn from Bernoulli distributions. For two attributes, the student's attribute profile indicators $a_{j1}$ and $a_{j2}$ were drawn from *Bernoulli*($\{1+\exp(-1.7\ (\theta^j - (-1))\}^{-1}$), and *Bernoulli*($\{1+\exp(-1.7(\theta^j - 1))\}^{-1}$), respectively. Then, $c_j = (a_{j1}, a_{j2})$ became the cognitive profile of student $j$. These student latent classes and item parameters ($s_i$, $g_i$) allowed the generation of response data of all the students for all items using the conditional probability computed by equation (2). Again, R was the software package used to simulate data (R Core Team, 2017).

### 3.2.2. Parameter Estimation and Data Analysis

Estimation and data analyses in this study were nearly identical to the procedure described in Study 1. Indeed, all three models were fit to the data simulated in Study 2 and followed the methods described in that study to investigate the ordering of learning levels and plausibility of combinations of levels. In Study 2, only two adjustments were made. The first revision was to use the known parameters of HO-DINA to generate response data to identify the true classification of students into combinations of levels. This classification became the baseline criteria to assess the performance of the models in recovering the true information contained in the data. The second change in Study 2 was the way the classification accuracy and cross-model classification consistency for two pairs of models is aggregated. In this study, the classification of students into learning levels by HO-DINA in the true scenario was assumed to be theoretically correct. Students were classified into four cognitive profiles [00], [10], [01], and [11] by fitting HO-DINA to the data. Since the [01] doesn't correspond to any of the three learning levels, simulated students classified into this profile were excluded from the analyses of classification accuracy and cross-model classification consistency. In the true scenario, the expected proportion of students in this profile was less than 2% when the attributes were of extreme distance, and about 10% when they were moderately distant. The computation for these percentages can be found in Appendix 1. Using the true level classification, the classification accuracy and cross-model classification consistency for combinations of levels between MIRT-SS, and DINA with HO-DINA were calculated for the true scenario. For brevity, the remaining details of parameter estimation and data analyses for this study will not be repeated. In the last study, all the models will be fit to three sets of empirical data collected to evaluate the learning progression theory described in section 2.1.2 of Chapter 2.

### 3.3. Study 3: An Empirical Application

In this empirical analysis, all the three models mentioned previously will be used to calibrate response data collected to evaluate the theory for LF and PR. Model-data fit was examined. Then, results were interpreted in reference to the predictions informed by the theory and findings of Study 1 and 2. In the next sections, key aspects of this study are described in detail.

### 3.3.1. Data

The empirical data contain item responses for 216 items by a sample of about 4,000 students from grades 6 through 8. Those items were developed to measure two learning progressions: LF and PR. There was a mixture of polytomous and dichotomous items in the item set. Among the items, 37.5% (81) were polytomously scored. Each progression was theorized to have five learning levels (Arieli-Attali et al., 2012). Due to testing time limitations, most students only answered about 20 items from both progressions, and each item had responses from approximately 400 students. Under this design, each student did not take items from all the four groups of items measuring the five learning levels. Typically, some items from two adjacent groups (e.g., level 1-2, level 2-3) were chosen to build test forms. Given that for each student there existed at least one item group from which the student did not answer any items, the DINA or HO-DINA models of four attributes defined by the four item groups could not be fit. However, if the data set was partitioned into three subsets containing response data for items from two adjacent item groups, the responses became suitable to fit both DINA and HO-DINA. For example, in Pham et al. (2017), the authors selected data for items from item group 1 and 2 that measured level-1 and 2, and level-2 and 3, respectively. In doing so, the data in Pham et al. (2017) had 589 students with responses to 72 LF and 48 PR items. Each student answered about 20 items, among which at least two items were from item group 1 and 2 of each progression. In this data set, each item had responses from 137 to 303 students. With those features, all the three models (i.e., MIRT-SS, DINA, and HO-DINA) could be fit to the data.

To extend previous work presented in Pham et al., (2017), three data sets from the original response matrix for all the 216 items of LF and PR were extracted. The first set is the data to which we fit DINA model in Pham et al., (2017). This data set is for items in the first two item groups (i.e., items measuring level-1 and 2, and level-2 and 3), and all students answering at least two items in each of the four item groups of two progressions were included. The second data set is for all items in the next two item groups (i.e., items measuring level-2 and 3, and level-3 and 4) of LF and PR, and all students who answered at least two items in each group are selected. Similarly, the last set is for the last two item groups (i.e., items measuring level-3 and 4, and level-4 and 5). Among the 10 postulated combinations of levels for LF and PR, the first five can be evaluated using the first data set. These combinations included (1, 1), (1, 2), (2, 2), (2, 3), and (3, 3). In those notations, the first and second indexes represent the learning levels of LF and PR, respectively. The next two combinations of (3, 4) and (4, 4) can be examined by the second data set. Finally, the last two (5, 4) and (5, 5) can be investigated using the last set of data.

In Pham et al. (2016), a graded response model was used to calibrate polytomous items. In this follow-up study, the response data was simplified by dichotomizing all the 81 polytomous items. The reasons for this treatment are twofold. First, it is slightly easier to work with only dichotomous data. Second, dichotomous items are still in wide use for learning progression-based assessments (e.g., Confrey et al., 2017; Shin et al., 2017). The first reason is especially relevant for DINA, and HO-DINA. As of this writing, the package used to fit those models (GDINA) only supports model-data fit investigation for dichotomous data (Ma & de la Torre, 2017). Equally important, the dichotomization of polytomous items in the empirical data also made it easier to use the findings from simulation studies 1 and 2 to interpret the empirical analysis results.

### 3.3.2. Parameter Estimation and Model-data Fit

Our prior experience working with these empirical data suggests that there might be some items with unreasonable parameter estimates in the first round of calibrations by MIRT-SS, DINA or HO-DINA. For instance, some items might have very large or negative discrimination parameter estimates when we fit MIRT-SS to the data. To mitigate the impact of those items, items with discrimination estimates falling out of the range [.25, 3.5] in MIRT-SS final calibrations were excluded, as was done in Pham et al. (2016). Similarly, item discrimination indices obtained from fitting DINA and HO-DINA to the data should not be negative (Lee et al., 2012). After normal convergence in the final calibrations was achieved, model-data fit could be evaluated; how fit was assessed will be discussed in the following paragraphs.

For MIRT-SS, since this model is equivalent to a structural equation model (SEM) of two correlated factors (Takane & de Leeuw, 1987), some SEM-based fit indexes provided by flexMIRT (Cai, 2015) were examined. Those indexes included root mean square of error of approximation (RMSEA), and Tucker-Lewis index (TLI). At the item level, standardized LD $X^2$ for each item pair (e.g., Chen & Thissen, 1997) was collected from flexMIRT calibrations to investigate model-data fit. For DINA, three absolute fit statistics at item level including the proportion correct, transformed correlation, and log-odd ratios were tabulated for each converged calibration (Chen, de la Torre, & Zhang, 2013). Then, using Dunn-Bonferoni correction to evaluate model fit for the items under the CDM framework, the maximum z-score tests for each of the three statistics were conducted. To compare the relative model fit of the DINA and HO-DINA models, three statistics which included deviance, AIC, and BIC (Chen et al., 2013) were used.

### 3.3.3. Data Analysis

Analysis of the statistical results obtained from fitting the three models to the data in this empirical study was very similar to what has been described for Study 1. In comparison to the simulation-based investigations, only one amendment to this exploration was adopted. As a matter of fact, one doesn’t know which of the models can explain the empirical data perfectly. Thus, it was not possible to compute the classification accuracy of sorting students into combinations of levels since the true classification in the empirical data was unknown. Instead, the cross-model classification consistency of each pair of models was computed. The results obtained from fitting the models to the data sets were interpreted in reference to the 10 postulated combinations of levels described in Table 2.1.2 in Chapter 2, the findings from Studies 1 and 2, and published works on this topic found in the literature.

### 3.4. Summary of Research Method

In this chapter, three studies to investigate the effectiveness of MIRT-SS, HO-DINA, and DINA in analyzing assessment data to evaluate learning progression theories were described. In the simulation studies (i.e., Study 1 & 2), the effectiveness of the models was examined similarly across studies by looking at how well they recovered the true information of the simulated learning progressions contained in the generated data. For the empirical analysis, model-data fit and evidence regarding the ordering of LF and PR items and the plausibility of level links were the main sources of information to assess the effectiveness of the model. Some practical implications of this study can be seen via several lenses. From a modeling perspective, this study was expected to help us understand how the MIRT-SS, HO-DINA, and DINA would behave in the best-case scenarios in which one knows the trustworthiness of the underlying learning theories. It was the first time a study of learning progressions using both IRT and CDM models involved a simulation component to inform the interpretation of empirical findings. Through Studies 1 and 2, an understanding of the models was strengthened. As we moved to analyze the empirical data, what was learned from the simulation would allow us to make informed conclusions about the theory of LF and PR using statistical evidence obtained from fitting the models to the data and knowledge of how they would behave in certain circumstances. From an application perspective, this study could be the first initiative to cast light on the effectiveness of using HO-DINA and DINA in building formative assessments using learning theories. If these CDMs are useful at evaluating learning progression theories and calibrating item banks, the next step would be using them to propose and investigate different design features used to build assessments for learning. From a scholastic point of view, Rasch model and thus Wright maps have been the main toolkit for researchers to empirically evaluate learning progressions. The tool seemed to serve the purpose quite well. However, a study of other models in this context is needed to empower more thorough investigations of learning progressions and suggest different modeling tools to build assessment instruments. The consideration of MIRT-SS, HO-DINA, and DINA in this study, which are different from the Rasch tradition, helps contribute to the literature of learning progressions and practical solutions available for building learning assessments.

### 3.5. Tables and Figures for Chapter 3

**Table 3.1.1. Q-matrix for One Learning Progression**

| | **Attribute-1**<br>(defined by item group 1) | **Attribute-2**<br>(defined by item group 2) |
|---|---|---|
| Item1.1 | 1 | 0 |
| … | 1 | 0 |
| Item 1.$n$-$_{1-2}$ | 1 | 0 |
| Item 2.1 | 0 | 1 |
| … | 0 | 1 |
| Item 2.$n_{2-3}$ | 0 | 1 |

**Table 3.1.3. Summary of Study 1**

| Condition | Scenario | Generating Parameters | | | | Fitting Models | Data Analysis |
|---|---|---|---|---|---|---|---|
| | | *N* | *I* | ***θ*** | *ρ*($\theta_1$, $\theta_2$) | | |
| 1 | True | 500 | 40 | $\mathcal{BVN}(0,1)$ | .6 | MIRT-SS, DINA, HO-DINA | Comparison of IRT item difficulty; Attributes' hierarchy by locations under HO-DINA and percentages of students in the inconsistent profile; |
| 2 | | | 40 | | .9 | | |
| 3 | | | 60 | | .6, | | |
| 4 | | | 60 | | .9 | | |
| 5 | | 1,000 | 40 | | .6 | | |
| 6 | | | 40 | | .9 | | |
| 7 | | | 60 | | .6 | | |
| 8 | | | 60 | | .9 | | |
| 9 | False | 500 | 40 | | .6 | | Classification accuracy and Cross-model classification consistency and by DINA vs. MIRT-SS, and HO-DINA vs. MIRT-SS |
| 10 | | | 40 | | .9 | | |
| 11 | | | 60 | | .6 | | |
| 12 | | | 60 | | .9 | | |
| 13 | | 1,000 | 40 | | .6 | | |
| 14 | | | 40 | | .9 | | |
| 15 | | | 60 | | .6 | | |
| 16 | | | 60 | | .9 | | |

# CHAPTER IV

# RESULTS

In this chapter, the results of the simulation and empirical analyses will be presented, and the findings will be summarized to inform several conclusions presented in Chapter 5. All three studies aim to investigate the effectiveness of MIRT-SS, HO-DINA and DINA in evaluating two claims of learning progressions considered in this dissertation. The first claim is about the correct ordering of learning levels within each progression. The second claim deals with the co-occurrence of learning levels across progressions. Given the commonality among the studies, results will be reported for each claim of each study in the following sections and recapped at the end of the chapter. First, the contents are organized by study (i.e., studies 1, 2 & 3). Second, within each study, results of the five methods examining the order of learning levels will be shown. Then, after the order is established, evidence evaluating the second claim of level links will follow. Tables and figures are numbered by the studies and displayed towards the end of the chapter. Finally, the chapter will conclude with a summary of the findings across studies.

## 4.1 Study 1: MIRT-SS as the Generating Model

In this study, 16 conditions under the true and false scenarios were considered using MIRT-SS to simulate response data for two learning progressions. To account for sampling error, 100 replications for each condition were conducted. Regarding the true scenario, the first eight conditions reflected valid progressions in which their learning levels were ordered from low to high with respect to difficulty. Regarding the false condition, item difficulties of items measuring different learning levels were sampled from the same standard normal distribution indicating that the levels were not correctly ordered. To evaluate how the models addressed the second claim related to level links, the classification of simulated students into combinations of levels using the generated item and proficiency parameters under the MIRT-SS framework was treated as the true classification. Then, the classification resulted from fitting MIRT-SS, HO-DINA, and DINA to the simulated data was compared with the true one to cast a light on how well each model recovered the true categorization using the generated parameters. The results to address each aspect of the theory underlying the simulated progressions from each of the models considered in this study will be then be presented. Tables shown toward the end of this chapter contain the details of the results. For some cases, plots were created using the statistics from the tables to help better visualize the results.

### 4.1.1. Claim 1: Ordering of Learning Levels within Each Progression

As laid out in Section 3.1.3 of the Method chapter, five methods were adopted to analyze learning progressions data to evaluate the first claim of level ordering. The methods included (i) One-tailed t-tests of item difficulty, (ii) Order-tests of cut scores for MIRT-SS, (iii) Attribute location tests, (iv) Inconsistent profile minimum tests for HO-DINA, and (v) Inconsistent profile minimum tests for DINA. Among the methods, only the t-tests of item difficulty were a hypothesis testing procedure. The remaining four approaches were based on binary classification tests. Given the nature of the tests, the term "*true positive*" will be used to indicate cases in which the correct order of learning levels was present (i.e., true), and the tests confirmed that information (i.e., positive). Similarly, the term "*false positive*" will be used to signify that the correct order was absent (i.e., false), but the tests indicated that the increasing order existed (i.e., positive). When sampling error is taken into consideration, the rate of false positive replications becomes type I error rate, the true positive rate is the same as statistical power in hypothesis testing. Methods that result in high true positive and reasonable false positive rates will be considered superior to their counterparts with low true positive and less reasonable false positive rates.

True and false positive rates for the five methods validating the ordering of learning levels are reported in Table 4.1.1. Several themes are observed from the table. First, none of the methods seemed to perform dominantly better than the others in inspecting level ordering. On the one hand, false positive rates for the MIRT-based t-tests method appeared to be very small. For all the eight false conditions, there were less than two cases out of 100 replications in which the one-tailed t-tests falsely rejected the null hypothesis of incorrect level ordering for both progressions. On the other hand, this test is likely to be overly-strict due to its low power as observed in the true conditions. The true positive rates for this test ranged from 41% to 80%. The power rates increased with larger sample sizes and longer tests. Power did not seem to be influenced by the correlation between progressions. In short, the MIRT *t*-tests appeared to be a very powerful tool to detect fallible cases. Whereas, its ability to validate true level ordering was questionable due to its low to moderate true negative rates. The lack of power of this test can be explained by the quite small number of items simulated in this study. If more items per item groups were generated, the power of the MIRT t-tests could have been higher.

Second, the results obtained from the other four methods were contrary to those using MIRT t-tests. In effect, their false positive rates varied from 15% for the HO-DINA location test of condition 16 to 41% for the HO-DINA minimum test for condition 13. Those very high error rates signified that the four tests lack the power to detect incorrect level ordering of the progressions. On the optimistic side, those tests appeared to obtain promising true positive rates. Especially, the MIRT order-test had all the rates greater or equal to 95%. When one has up to 1,000 students and 60 items for four item groups, it was very likely that the MIRT order-test would correctly reconfirm the ordering of learning levels for all the replications. For other methods, true positive rates tended to be lower and varied from 98% for DINA minimum test in condition 8 down to 73% for the HO-DINA minimum test in condition 1 (i.e., 500 students and 40 items). Across conditions and methods, it is expected to see that the true positive rates were likely to increase with more students and items. In terms of failing to reject the null, the connection between false positive rates and sample sizes was not clear for the MIRT t-tests and order-test. However, it was quite notable that the false positive rates of the tests based on HO-DINA and DINA calibrations tended to decrease with sample sizes. The lowest rate of 0.15 occurred for the HO-DINA location test in condition 16 in which one had the maximum of students and items considered in this study. The highest false positive rate was around .40 which was observed for the CDM-based methods in conditions of either fewer students or items. Due to similarity of results across conditions of the same sample sizes and correlation between two progressions, two plots were created to visualize the results explained above for conditions 1 and 8. They can be found in Figures 4.1.1 and 4.1.2 at the end of this chapter. In the next section, the results using the methods to assess the second claim of level links will be reported.

### 4.1.2. Claim 2: Co-occurrence of Learning Levels across Progressions

As discussed in Chapter 3, using MIRT-SS, HO-DINA, and DINA, one can classify simulated students into learning level-1, level-2 or level-3 for each progression. In this study, MIRT-SS was used to generate the response data. As a result, the known parameters for this model were adopted to determine the simulated respondents' "*true classification*". This classification, which was based on the true item and person parameters, was used as the baseline information to evaluate the classification for each the model. To be able to use MIRT-SS parameter estimates to classify students into learning levels, it is required that the cut scores to distinguish levels are at least in the correct order. Given that requirement, only the true scenario was considered and replications where the cuts for each progression were increasingly ordered were taken into account. The numbers of such cases for each condition can be found in the column “MIRT: Order-test” of Table 4.1.1. In what follows, the term “*correctly-ordered replications*” will be used to codify these cases.

To investigate the usefulness of the models in recovering the true classification into combinations of levels, classification accuracy rates for the models were averaged across correctly-ordered replications in each condition of the true scenario. In addition, the cross-model consistency of classifying students into level combinations between pairs of models was also aggregated across these cases. The higher the rate of classification accuracy, the better the model was at recovering the true classification. In a similar vein, the higher the cross-model classification consistency, the more similar the level classification by models were.

Table 4.1.2 reports the true proportions of students in each of the nine combinations of levels. The proportions were averaged across all 100 replications for each of the eight conditions in the true scenario, since the cuts for all of them were in increasing order. The false conditions for the first claim of level ordering were not considered, since the claim was unlikely to be supported in this scenario. As shown in Table 4.1.2, more students were placed into the same levels across progressions than into combinations of different levels. For example, the averaged percentages of students in combinations [11], [22] and [33] varied from 15% to 23%. Meanwhile, those values for other level links were from 0 to 11%. It is also seen quite clearly that the percentages in the combinations of the same levels were higher when the correlation between progressions was stronger (conditions 2, 4, 6, and 8). This result is reasonable, since in this case student scores across progressions tend to be more similar than in the case in which the correlation is weaker.

Results from fitting MIRT-SS for the correctly-ordered replications in this study are displayed in Table 4.1.3. Comparing these results with the true proportions shown in the previous table, one can see that MIRT-SS suitably recovered the true proportions of students in the nine level combinations. In all conditions, the differences between the estimated proportions by MIRT-SS and the true ones were quite minimal. For example, about 40% of the 72 proportions for nine combinations across eight conditions in Tables 4.1.2 and 4.1.3 were identical in values. Most differences in the remaining cells were within .01 to .02. This finding is an indication that MIRT-SS seemed to recover the true classification well.

As for HO-DINA and DINA, Table 4.1.4 contains the proportion of students classified in each combination of levels. Similar to the case of MIRT-SS, the proportions were averaged across correctly-ordered cases for the conditions in the true scenario. A few patterns emerged from the results. First, the classification into level combinations by the CDMs was nearly identical within each condition. The differences of the proportions of students classified in each combination by HO-DINA and DINA were within .00 to .03 for all the cells. Second, the classification by these models appeared to be quite different from that of MIRT-SS. Indeed, across true conditions, more than half of the students were classified into two combinations of level-1 and level-3 for both progressions (i.e., combinations [11] and [33]). The proportions for these two combinations by the CDMs seemed to be much higher than those by MIRT-SS. On average, about a third of the students were classified by either HO-DINA or DINA into each of the combinations. Meanwhile, the proportions by MIRT-SS were around .20. Second, another notable difference was that the proportions of students who were in level-2 of both progressions varied from .15 to .25 for MIRT-SS across true conditions. However, DINA consistently classified none of the students into this combination across all the conditions. Similarly, the proportions of students in level-2 for each progression by HO-DINA stayed as minimal as .01 throughout all the true conditions. Third, differences in classification results, using each of the three models, can also be seen in other combinations of levels. The CDMs tended to locate more students into combinations [13] and [31]. In contrast, MIRT-SS seemed to yield more students in the level links of [12], [21], [23] or [32]. Last, as can be seen in Table 4.1.5, HO-DINA and DINA located some simulated students into the inconsistent profile [01] of one progression. In the table, letter "I" was used in the first or second place to indicate that the CDMs classified some students in profile [01] for the first or second progression. For example, combination denoted by "1I" contains students classified in level-1 of the first progression and inconsistent profile of the second progression. Across the true scenario and on average over correctly-ordered replications, no more than 1% of the students were placed in each of the seven combinations of at least one inconsistent profile. Notably, DINA classified no students into any of the inconsistent combinations. Equally remarkable was that none of the models located any students in the inconsistent profile for both progressions. These results support the effectiveness of the CDMs in analyzing data of valid learning progressions. To aid in interpretation, Figure 4.1.3 visualized the results of how each model enabled us to classify students into combinations of levels for conditions 1 and 8 of this study. Plots for other conditions of the moderate and high correlation were similar to the first and second panels of the figure, respectively. In summation, the performance of MIRT-SS, HO-DINA, and DINA in identifying student learning profiles in this study was shown to be quite different. These discrepancies among the models will be further discussed in the final chapter.

The last set of analyses done in this study collected the classification accuracy and cross-model classification consistency of the models when they were used to classify students into combinations of levels in the true scenario. To examine the impact of using significant or correctly ordered cases on the rates, two sets of statistics were computed for each condition. The first one was the accuracy and cross-model classification consistency rates across all correctly ordered replications. The second one is for all replications with significant t-tests results comparing item difficulty for both progressions. Table 4.1.6 aggregated the accuracy rates for each condition and method (i.e., ordered or significant cases). The results reveal several themes. On the one hand, the accuracy rates in the best-case scenario when both MIRT-SS was used to generate and analyze the data ranged from 60% to 69%. The rates did not seem to be dependent on the nature of the replications (i.e., correctly ordered or significant). On the other hand, the accuracy rates for HO-DINA and DINA were lowest at 33% and highest at 45% which were smaller than those by MIRT-SS. Moreover, the rates by the CDMs appeared to be higher when all the correctly-ordered cases but not only the significant ones were taken into the computation. Given that the significant cases were a subset of the correctly ordered ones, it can be implied that on average, the accuracy rates for the insignificant correctly ordered replications seemed to be slightly higher than those for the significant ones.

Table 4.1.7 contained the cross-model classification consistency rates between HO-DINA, DINA and MIRT-SS when different sets of replications were considered. Overall, the rates were very similar across conditions and varied from .35 to .49. The consistency between HO-DINA and MIRT-SS seemed to be slightly larger than between DINA and the baseline model. This result can be attributed to the unidimensional structure underlying each progression of HO-DINA and MIRT-SS. Similar to the classification accuracy, the cross-model classification consistency rates appeared to be higher when all the correctly ordered replications (not only the significant cases) were included in the computation of the rates. The differences between the rates of the two methods were about .03 to .10. In short, the classification accuracy and cross-model classification consistency for the models in this study were far from perfect. For the best case, when MIRT-SS was both used to generate and analyze the data, the highest accuracy rate stayed at 69%. The lowest rate was of 33% for the classification of DINA and the true classification. These far-from-perfect rates indicate that identifying student learning profiles is challenging when one only relies on statistical models and methods, since there are many ways the classification could go wrong. This point will be further discussed in the last chapter of this dissertation. In the next sections, the results for Study 2 will be shown, following the same content structure for the first study. Since two sets of cases were simulated within the true scenario, findings for each set will be presented sequentially within the sections for each of the claims that follow.

### 4.2. Study 2: HO-DINA as the Generating Model

In Study 2, the generating model was switched from MIRT-SS to HO-DINA. To examine how well the models considered in this dissertation detect the ordering of the levels, two sets of data within the true scenario were simulated. For the extreme difference cases, the two attributes defined by two item groups were generated to be extremely different, locating the first attribute at -1 and the second one at 1 in the logit scale. Thus, the distance between the two attributes was 2 logits in these cases. In the moderate difference replications, the location of attribute associated with levels 1-2 was fixed at -.25 and that of the higher levels was set at .25. As a result, the location distance between two attributes within one progression was only .5 rather than 2 in the second set of data. It is expected that the true positive rates for the conditions of distant attributes will be greater, and the false positive rates for these cases will be smaller than those of the closer attributes. In addition, the classification accuracy and cross-model classification consistency measures are also expected to be higher in the former than in the latter. The results for these investigations are reported next. Following the structure of Section 4.1, the information will be organized by two claims of the underlying theory in each scenario.  Complete results will be shown in tables, and some selected plots will be created to highlight the results.

### 4.2.1. Claim 1: Ordering of Learning Levels within Each Progression

#### 4.2.1.1. Ordering of Levels for the Case of Extreme Difference

When the five approaches to detect the ordering of learning levels were carried out as used in Study 1, it was found that all of them yielded expected results for all 100 replications for cases of extreme difference (i.e., attribute location distance equals two logits) in the true scenario. In other words, two MIRT-based methods (one-tailed t-tests of item difficulty, and tests of ordered cuts) and three CDM-based approaches (HO-location test, HO and DINA-minimum tests) revealed positive results that support the correct order of learning levels for every replicated data set in these conditions.  Table 4.2.1 displays the true and false positive rates for the extreme difference cases. It is seen that the true positive rate was 1 for all conditions in the true scenario, which mean that all five methods confirmed the true information used to generate the data.

For the false scenario, the result pattern was similar to the finding of Study 1. In effect, false positive rates for the MIRT-based t-tests of item difficulty were very small. Half of the rates were zero and all of them were below .03. In other words, the t-test worked almost completely to detect the incorrect order of learning levels across replications in the false scenario. These low false positive rates of the MIRT-based t-test signifies the usefulness of the method in investigating the order of learning levels. If the test reveals an insignificant result comparing item difficulty of two adjacent item groups, it is likely that knowledge and skills measured by items from these groups are not in the right order of increasing complexity.

Regarding the false positive rates indicated by the remaining methods, it was observed, as in the case of Study 1, that these approaches were not very effective at reconfirming the true information in the simulated data within the false scenario. Indeed, the test of ordered cuts, location test for HO-DINA, and minimum tests for HO-DINA and DINA rejected incorrect order for 19% to 35% of the cases in each condition. The error rates for the first two tests were similar. Whereas, false positive rates of the minimum test using profile proportions by HO-DINA and DINA were nearly identical and consistently higher than those of the first two tests. Again, these results appeared to be in line with the findings in Study 1 for the methods.

#### 4.2.1.2. Ordering of Levels for the Case of Moderate Difference

When the magnitude of the difference between attribute locations was reduced from 2 to .5 logits, the impact of the reduction can be partially seen in Table 4.2.2. On the one hand, the CDM-based tests made no incorrect detections in the true scenario as in the case of extreme difference between attributes. In fact, the methods successfully reconfirmed the correct order of the learning levels for 100% of the replicated data sets in the true scenario. This result is explainable, given that HO-DINA was used to generate data in this study and the classification of students into cognitive profiles by HO-DINA and DINA should be close to identical. On the other hand, the two MIRT-based tests showed a notable drop in true positive rates when the difference between attributes was moderate. Indeed, while the rate for the test of ordered cuts was reasonable, the results for the t-tests of item difficulty differences reached the highest true positive rate at 78% for one condition and became lowest at 49% for the other. These moderate to low true positive rates were the most remarkable difference, casting a light on the sensitivity of MIRT-SS in detecting the ordering of levels when the distinction between them varied from moderate to extremely large. The MIRT-based methods tended to be more sensitive to the magnitude of the difference between item groups than the CDMs. One can also deduce that the CDMs can recover the ordering of learning levels very well, if these models fit perfectly with the data.

### 4.2.2. Claim 2: Co-occurrence of Learning Levels across Progressions

In this study, HO-DINA was used to generate data. Then, MIRT-SS, HO-DINA and DINA were fitted to determine how well the models recover the true classification of students into learning levels by the higher order CDM. As stated earlier, two sets of data were simulated for each of the true conditions. The results for the classification accuracy and cross-model classification consistency for each set of data will be reported in the following paragraphs.

#### 4.2.2.1. Co-occurrence of Levels for the Case of Extreme Attribute Difference

For the extreme difference case, Table 4.2.3 contains the classification accuracy for three models with respect to the true classification and cross-model classification consistency for MIRT-SS and DINA in comparison to HO-DINA as the generating model. It is observed from the table that the two CDMs seemed to perfectly reproduce the true classification. The lowest accuracy rate for these models was as high as 97% for the conditions with fewer items (i.e., 40). When a condition had 60 items, the accuracy rate went up to 99% for both HO-DINA and DINA. Since HO-DINA was used to generate data, it is expected that the accuracy rates for MIRT-SS with the true classification would be smaller than the CDMs’. The rates for the MIRT model seemed to reflect the expectation. They varied from 81% to 83% for conditions of 60 items. And, the rates were slightly smaller at 75% or 76% when one had fewer items (i.e., 40). Given that the

CDMs almost perfectly reproduced the true classification, the cross-model classification consistency between MIRT-SS and HO-DINA was almost the same as the classification accuracy of the former. The last column of table 4.2.3 shows that the classification by fitting the CDMs to the simulated data was completely identical. This result can be explained by the fact that IRT-parameters for the attributes of HO-DINA were estimated after the cognitive profiles of each student were identified by the DINA (J. de la Torre, personal communication, April 30, 2017).

Since the classification accuracy and cross-model classification consistency in the extreme difference case were high, the average proportion of students classified in each combination of levels by the models in the true scenario should look similar. Indeed, Table 4.2.4 displays the proportions averaged across all 100 replications of the students being categorized into nine reasonable combinations. Three sets of findings seemed to emerge in this case. First, the values in each cell of the table signified that classification results by the CDMs into level combinations appeared to be nearly identical with the true one for almost all the true conditions. Among all, a difference of .01 between the proportions by HO-DINA and DINA and the true one only occurred in three conditions and combinations. They include (i) combination [12] of condition of 1,000 students, 60 items and moderate correlation, (ii) combination [31] of condition of 1,000 students, 40 items and moderate correlation, and (iii) combination [32] of condition of 1,000 students, 60 items and strong correlation. For all the other conditions and combinations, the averaged proportions for each combination by the CDMs were the same as the ones computed using the true parameters. Second, the difference of MIRT-based proportions and the true one by HO-DINA was more salient. Nonetheless, the magnitude of the differences was consistently within the range of .01 to .03 across all the conditions. Third, about one third of the simulated students were identified into level-2 of both progressions (i.e., level combination [22]). This pattern was consistently observed in all the models across all true conditions. This finding is reasonable given that the distance between two attributes in this case was very wide (i.e., 2 logits). The extreme distinction between attributes within each progression simulated in the HO-DINA was very likely to result in distant cut scores to distinguish levels 1-2 and levels 2-3. When the cuts were far away, more students would have been placed into the middle level (i.e., level-2) leading to the large proportions of students in combination [22]. For illustrative purpose, Figure 4.2.1 visualizes the proportions of students classified in nine reasonable combinations of levels for conditions 1 and 8 of this study.

As mentioned earlier, the CDMs can locate students in inconsistent profile of [01]. Students in this latent class master knowledge and skills of higher but not the lower levels. Students can be placed in the inconsistent profile for one or both progressions. Consequently, there were seven level combinations that contain at least one inconsistent profile.  Just as in Study 1, the letter “I” was used to indicate the inconsistent combination. Table 4.2.5 reports the proportion of students classified into the inconsistent combinations.  One notable theme can be seen from the table. That is the consistency of results across conditions and models. On average and for all the true conditions, there were 1% of students classified in combinations “2I” and “I2”. The proportion for the five remaining combinations was all 0. It is noted that due to rounding error, the total proportion for the True, HO-DINA and DINA classification in Table 4.2.5 did not add up with the respected total proportion in Table 4.2.4 to 1. In the next case where the distance between attributes was simulated at .5, it is expected that the classification results obtained from fitting the models will show more students in the inconsistent combinations.

#### 4.2.2.2. Co-occurrence of Levels for the Case of Moderate Attribute Difference

As expected, the classification accuracy of MIRT-SS for data generated with shorter distance between learning levels was lower than results of the previous case reported earlier. Table 4.2.6 reports the average classification accuracy and cross-model classification consistency for the models when replications with significant results for the t-tests of item difficulty were taken into the computation. Similarly, table 4.2.7 contains the results when all replications whose MIRT-based cut scores distinguishing adjacent levels were correctly ordered. As a reminder, the proportions for these significant and correctly-ordered replications can be found in Table 4.2.2.

The accuracy and cross-model classification consistency rates in this case imply two key themes.  First, as reported earlier, the CDMs seemed to recover the true classification generated by HO-DINA extremely well. When all the correctly ordered replications were included in the analysis, the models correctly reproduced 97% and 99% of the true student learning profiles for assessments of 40 and 60 items, respectively. These results were the same as the accuracy percentages of the models when the distance between attributes was extreme. Once again, classification of HO-DINA and DINA were identical across all conditions. Second, classification accuracy of MIRT-SS and the cross-model consistency of this model and HO-DINA became much lower in this case in comparison to the case of extreme discrepancy. Indeed, the average across correctly-ordered replications of the accuracy and consistency rates of MIRT-SS, in this moderate difference case, center around .49 to .52. The rates for assessments with more items were slightly higher than those with fewer items. However, the difference was very minimal. The lower rates for the moderate distance attribute conditions using MIRT-SS suggest that the cross-model classification consistency between this model and HO-DINA is dependent upon the magnitude of the discrepancy between learning levels. If the knowledge and skills exhibited by the levels are more distinguishable, it is more likely that the IRT and CDM models will yield more highly consistent classification and vice versa.

To report the results of how the models categorized students into combinations of levels in this case, tables 4.2.8 and 4.2.9 tabulate average proportions of students identified in the nine possible combinations of levels, and seven combinations of at least one inconsistent profile, respectively. Comparing these tables to that for the extreme difference case, three lines of findings emerge. First, the results with MIRT-SS appeared to be much more distinct than those using CDMs. This statement can be supported by the high proportions of students classified in combination [11] by the MIRT-SS and the very low proportions for level links [22] by this model. Consistently seen across conditions, MIRT-SS identified about one third of the students into combination [11]. The proportions of levels linked by the CDMs were around one fifth or less. Second, due to the moderate distance simulated for the attributes, far fewer students were found in combination [22] in this case. This result holds true across all models and conditions. Indeed, MIRT-SS only placed 1% to 2% of the students in this level link. Whereas, the percentages using CDMs were higher but stayed around 6% to 7%. These figures from HO-DINA and DINA were very close to the true percentages generated by the model. It is noted that combination [22] contained about one third of the students in the extreme attribute difference case. To aid graphical interpretation, Figure 4.2.2 visualizes the proportions of students in nine reasonable combinations when the difference of the attributes was moderate. Again, the figure can help us see clearly that the classifications of the MIRT model were more distinguishable from those of the CDMs, in this moderate condition, than in the extreme difference case. Last, there were many more students classified into inconsistent profiles in this case than in the previous one. On average, approximately 20% of the students were placed by the CDMs into at least one inconsistent learning profile. This value is much higher than the average of approximately 2% of students classified in at least one inconsistent profile in the previous case (see Table 4.2.5). The three sets of afore-mentioned findings signify that the cross-model classification consistency between the MIRT-SS and the CDMs depends on the magnitude of the distinction among the learning levels. The more distinguishable they are, the more consistent the classification results by the models becomes. Similarly, the magnitude of the difference was also reflected in the percentage of students classified by the CDMs into inconsistent profiles. The more different the levels, the less likely it will be for a student to receive an inconsistent classification.

### 4.3. Study 3: An Empirical Application

In this empirical study, all the three models considered in this dissertation were fit to three data sets of two learning progressions: LF and PR. Originally, the progressions were theorized to have five levels (Arieli-Attali et al., 2012). Previously, Pham et al., (2016) used a MIRT model to evaluate the theory and was able to support almost all aspects of the theory for LF and PR. As explained earlier in Chapter 3, to be able to fit the CDMs to reevaluate the theory, the whole response matrix for LF and PR was partitioned into three data sets. The first data set was for the first three levels of these progressions. The next one contained items measuring levels-2, 3 and 4. Finally, the last one was subset from the master data set for LF and PR to comprise students' responses for items of levels-3 to 5. This investigation, then, can be viewed as an application using the models to reevaluate the progressions. In the following sections, the results of this study will be reported. In the last chapter, these findings will be discussed and interpreted in light of the theory underlying the progressions and the results of Studies 1 and 2.

### 4.3.1. Calibration Results and Model-data Fit

#### 4.3.1.1. Item Exclusion

MIRT-SS model was fit to each of the three data sets. Items with extreme discrimination estimates (e.g., negative or in the two-digit numbers) or very large standard errors for any of the parameters were excluded in the next round of calibration. After several rounds of item removal using the criteria described in the method section, 6, 9 and 12 items were excluded from the first, second and third data sets. Estimates of item parameters from final successfully-converged calibrations were collected for further analyses. Overall, discrimination estimates of items in all the three data sets varied from .11 to 4.7. The mean and standard deviation of those estimates were 1.61 and .77, respectively. A few items with positive discrimination estimates smaller than .25 were retained due to the exploratory nature of this study. Thus, as many items as possible were kept to maintain a wider choice of items for follow-up data collections and/or studies using the items investigated in this study. Once the final sets of items were determined for each data set and item and student parameters collected from the final calibration using, flexMIRT, for the MIRT-SS model, and GDINA, for the CDMs, were used to evaluate the theory.

#### 4.3.1.2. Model-data Fit

When statistical models are used to analyze empirical data, it is a standard practice that one should check model-data fit before interpreting results (Swaminathan, Hambleton & Rogers, 2007). In this study, that procedure was followed by collecting as much information about model-data fit for the data sets as possible. Given the large amount of missing data by design, it was challenging to compute and aggregate all available fit statistics and indexes to paint a thorough picture of how the selected models fit with the empirical data. However, at least one global fit measure and one fit statistic at item level for each data set were tabulated. In what follows, model fit information will be presented.

Under the MIRT framework, the MIRT-SS model was fit to three sets of data (i.e., LFPR1223, LFPR2334, and LFPR3445). Due to the significant amount of missing data and large number of items in each data set, flexMIRT (Cai, 2015) was not able to output full-information global fit statistics. For all the calibrations, upon convergence, the program showed a note "*The contingency table is too large to compute the general multinomial goodness of fit statistics*" under the result section of "*Full-information fit statistics of the fitted model*". On the positive side, flexMIRT was able to compute the limited-information fit statistic $M_2$ (Maydeu-Olivares & Joe, 2005) and indexes based on that statistic. Values for the fit measures can be found in Table 4.3.1 at the end of this chapter. The tests of global fit using the $M_2$ statistic for all three data sets revealed significant results at alpha level of .05., with all p-values close to 0. Given the large sample sizes in each data set, it is expected that the results these tests of global fit would be significant. In this case, it is usually helpful to use other goodness-of-fit indexes to assess model-data fit. Columns 6 and 7 of Table 4.3.1 contain these indexes. RMSEAs of all the data sets were .03. The TLI varied from .85 for LFPR1223, to .90 for LFPR3445 and .92 for LFPR2334. These values are indications of "Close fit" between MIRT-SS and the data.

At item level, the standardized Chen-Thissen LD $X^2$ (Chen & Thissen, 1997) statistic was collected to examine the degree of fit of MIRT-SS for each pair of items. This statistic reflects how well the model explains the observed correlation of pairs of items. To put it another way, it examines the bivariate relationship or local dependency for each item pair. The model closely captures the correlational relationship for two items if the standardized LD $X^2$ for the pair is within the range of [-3, 3] (Chen & Thissen, 1997). The percentages of item pairs for which the statistics can be computed that met the fitting criterion above were shown Table 4.3.2. The results signified that MIRT-SS appeared to explain the bivariate relationship of the data well. Across all data sets, standardized LD $X^2$ statistics for more than 90% of the item pairs were not smaller than -3 or larger than 3. In summary, both global fit indexes and item-level fit statistics for MIRT-SS indicated close fit between the model and the data. Thus, the model appeared to explain the data well and parameter estimates can be used to evaluate the theory underlying the data. In the next sections, model-data fit for HO-DINA and DINA will be discussed.

For the CDMs, both relative and absolute fit statistics for the models were collected. As explained in the method chapter, AIC and BIC were used to compare HO-DINA and DINA. Table 4.3.3 reports these statistics for each data set. It is notable that the statistics were in favor of DINA across the calibrations. This model had one parameter less than its higher-order version and yet its AICs and BICs were consistently smaller than that of HO-DINA. All other factors remaining equal, it would be expected that the model with more parameters would exhibit better fit than one constrained to fewer parameters. For this reason, because it is a simpler model with better fit statistics, DINA is preferred to its higher-order version from a relative fit perspective.

At item level, two sets of item fit criteria were evaluated. The first set of criteria is the three statistical tests described in Chen, de la Torre and Zhang (2013): (i) the proportion correct test for each item, (ii) the transformed correlation test, and (iii) the log odds ratio test for pairs of items. Following the suggestion in that paper, the maximum z-score test using Bonferroni correction was used to eliminate the need to conduct many statistical tests for each item or item pair. This method also allows us to examine model-data fit at aggregated level across all items considered in this empirical study. The results of the maximum z-score tests for each data set are reported in Tables 4.3.4 and 4.3.5 for LF and PR items, respectively. The test statistics shown in the tables reveal three clear patterns. First, the results for HO-DINA and DINA across data sets were very similar. The maximum z-scores and p-values for the two models were nearly identical. The adjusted p-values by Bonferroni correction were somewhat different but led to the same conclusions of significance or non-significance for both models on all data sets. Second, the CDMs seemed to closely reproduce the observed proportion correct in the empirical data. Out of 12 maximum z-score tests, 11 of them produced an adjusted p-values greater than .05. The only case in which the proportion correct test revealed a significant result for both models was for data set LF2334, which was the set containing the largest number of items. It has 87 items in comparison with 73, 68, 49, 46 and 38 items for PR2334, LF1223, PR3445, PR1223, and LF3445, respectively. The larger number of items in LF2334 might increase the power of the maximum z-score test enough to detect the difference of model-implied and observed proportion correct across all the items in this data set. Third, the bivariate tests (i.e., transformed correlation and log odds ratio) yielded large test statistics and significant results for all six data sets. Both the p-values and adjusted p-values ones were consistently equal to 0 across all cases. The significance of the tests indicated that the models were very unlikely to sufficiently reproduce the empirical bivariate relationships for pairs of items. Given the large amount of data missing by design in these cases, it is understandable that the observed bivariate statistics for item pairs might not be reliable enough to be captured appropriately by the CDMs.

In the last effort to evaluate model fit for the CDMs, summary statistics of item parameter estimates for these models, following the recommendation by de la Torre and Douglas (2007) and de la Torre (2007), were gathered. The de la Torre and Douglas (2004) guidelines suggested that good-fitting items should have estimates for their guessing and slipping parameters smaller than .40. Otherwise, examinees, without mastering the attributes required by the items, can still experience unreasonable probability of endorsing said items. For item discrimination, de la Torre (2007) introduced a discrimination index $\delta_i = 1 - s_i - g_i$ for item $i$, where $s_i$ and $g_i$ are the slipping and guessing parameters for the item, respectively. This index reflects the magnitude of difference in the probability of answering item *i* correctly between an examinee who masters all the attributes required by the item and one who doesn't. The higher the index, the more discriminating an item. For this index, a value lower than .20 is considered to be low discrimination (Lee et al., 2012). Table 4.3.6 presented the proportions of items in each data set whose guessing and slipping parameter estimates satisfied the criterion of smaller than .40. It can be seen from the table that more than 70% of the items in every data set obtained a guessing parameter estimates satisfying the recommended indicator of good item fit. The mean values of these estimates varied from .20 to .26, and their standard deviations remained as low as .19 to .26 across all cases. The estimates for the slipping parameters were not as good as those for the guessing parameters. However, the majority of slipping estimates were within the range of [0, .40]. The percentages of these statistics that satisfied the .40 cutoff appeared to be higher for items measuring higher learning levels. In fact, 71% of LF3445 items had a slipping estimate below .40 for the CDMs. Whereas, only 56% of LF1223 items met this requirement. Along the same lines, it can be observed that items measuring higher learning levels had better item parameter estimates for the CDMs than their lower level counterparts. For both HO-DINA and DINA, the percentages of items in the data sets for higher levels that had better estimates tended to be higher than those for the data sets in lower levels. For example, 57% of LF1223 items obtained a slipping estimate smaller than .40, while the same percentage for LF3445 was slightly higher at 61%. There were a few exceptions for this result. For instance, the percentage of items with good slipping parameter for PR2334 was 10 points lower than that for PR1223. Table 4.3.7

showed the descriptive statistics for the discrimination indexes for each data set. First, the results for HO-DINA and DINA were almost identical. Although differences can be seen in some cells, they were all small and potentially negligible. Second, the summary statistics for the discrimination index indicated good fit at item level. Across the data sets, the means and standard deviations of the index varied from .34 to .49, and .16 to .20, respectively. Only one out of more than 200 LF and PR items had a negative discrimination index, which was very close to zero. The percentages of items whose discrimination indexes were larger than .20 ranged from 76% to 92% across cases. These results for the discriminating power of items under the CDM framework signified that this method seems to model the data well and can be used to classify examinees into cognitive profiles, thus learning levels.

Overall, relative fit statistics were in favor of DINA over HO-DINA. From the absolute and item fit perspective, both models seemed to fit moderately well with the data. Given the large amount of missing data and the exploratory nature of evaluating the underlying theory, the results from fitting the CDMs conclude these models can be used to examine the plausibility of learning progressions.

### 4.3.2. Claim 1: Ordering of Learning Levels within Each Progression

In this study, the ordering of learning levels was evaluated using difficulty estimates for items measuring different learning levels, under the MIRT framework, attribute locations, by HO-DINA, and proportions of students classified in the inconsistent profiles obtained from fitting HO-DINA, and DINA to the data. The ordering claim is supported if the difficulties of items measuring lower levels are lesser than those of items targeting higher levels. In the CDM framework, the theory will be defensible if locations of the attribute defined by knowledge and skills of lower levels are to the left of those defined by the higher levels. Similarly, if the proportion of students classified into the inconsistent profile is notably smaller than those of the profiles associated with the learning levels, the ordering of levels is supported. The following paragraphs report the results from fitting MIRT-SS, HO-DINA, and DINA to the data.

#### 4.3.1.2. MIRT-SS

Following the independent t-tests (one-tailed) method, item difficulty estimates of item groups measuring different learning levels were compared. Table 4.3.8 shows the results of those tests. As presented in the table, at a conventional alpha level of .05, five out of the six tests were significant with medium to large Cohen-d effect size measures. Indeed, items measuring levels 1-2 of LF (M = -1.17, SD = 1.74, n = 12) were significantly easier than items written to assess levels 2-3 (M = .76, SD = 1.61, n = 56) of this progression, *t*(15.3)= -3.54, *p* < .002, *Cohen-d* =1.19. The same statement can be made for items of levels 2-3 (M = .10, SD = 1.18, n = 58) and levels 3-4 (M = .39, SD = .91, n = 29) of LF, *t*(64.9)= -5.84, *p* < .001, *Cohen-d*=1.25. For PR, all three comparisons for this progression revealed significant differences. On average, PR items of levels 1-2 (M = -.03, SD = 1.3, n = 14) of this progression were easier than their peers from levels 2-3 (M = .79, SD = 1.20, n = 32), *t*(23)= -2.00, *p* =.03, *Cohen-d* = .66. Similarly, the mean item difficulty of levels 2-3 items estimated using data set LFPR2334 (M = .29, SD = .75, n = 32) of PR was statistically smaller than that of levels 3-4 (M = .81, SD = 1.01, n = 41), *t*(7.8)= -2.50, *p* = .01, *Cohen-d* = .57. Using the last data set (i.e., LFPR3445), it was observed that mean of difficulty estimates of items measuring PR levels 3-4 (M = .09, SD = 2.05, n = 40) were significantly smaller than that of levels 4-5 (M = .1.00, SD = 1.11, n = 9), *t*(23.1)= -1.82, *p* =.04, *Cohen-d* = .48. The only test with non-significant result was the one for LF items in levels 3-4 (M = .39, SD = .91, n = 32) and levels 4-5 (M = 1.01, SD = .82, n = 6), *t*(7.4) = -1.69, *p* = .07. This observation can be explained, in part, by the fact that the item group of level

4-5, from this progression in this data set, contained only six items.  This was the smallest number of items among all the item groups. The second method under the MIRT framework was to test the ordering of cut scores used to classify students into adjacent learning levels or, in short, the order-test. To facilitate the visual interpretation of this test, Figure 4.3.1 displays the boxplots of item difficulty estimates for six pairs of item groups. Two themes emerged from the figure. First, it can be seen from those plots that items measuring higher learning levels appeared to be more difficult for examinees than the ones targeting lower levels. Second, the medians of the item difficulties in the groups were all in increasing order, as one would expect. In other words, the order-tests returned positive results for all data sets, meaning we can use those medians to be the cut scores placing students in learning levels as explained in this dissertation's method section. It is noted that other methods to identify the cut scores, such as using test characteristic curves (TCC) and a suitable response probability (RP), are available (e.g., Hambleton & Pitoniak, 2006). In a previous study, Pham et al. (2016) used the median and the TCC methods with a RP of .50 and .66 to determine three sets of cut scores to evaluate the second claim of level links. It was found in that study that determining cuts by the three methods (i.e., using medians as cuts and using TCC with RP 50 and PR 66) yielded different results for each. Nonetheless, the proportions of students classified into combinations of levels using the different cuts were quite consistent and supported almost all predicted level links. This was the reason why only the median method was considered and used in this study.

#### 4.3.1.3. HO-DINA and DINA

The results of fitting HO-DINA to the data to evaluate the ordering of learning levels seemed to be in line with the evidence obtained from fitting the MIRT. Table 4.3.9 reported the locations of attributes-1 and -2, and the distance between them for each pair of item groups. For five out of the six comparisons, locations of attribute-1 were smaller than those of attribute-2. The only pair where the ordering of the location was not supported was LF levels 3-4 and 4-5 items. In this case, location of attribute-1 defined by knowledge and skills of LF levels 3-4 was slightly larger by .05 logits than the location of attribute-2 defined by the two highest levels (i.e., levels 4-5) of LF. This result is in accordance with the non-significant difference of item difficulties of LF items measuring levels 3-4 and 4-5 reported earlier. In summation, the HO-DINA locator tests revealed confirmative results supporting the ordering of learning levels for five out of six data sets with the only exception being LF3445.

Under the CDM framework, the plausibility of level ordering can also be evaluated by adopting the minimum test for the proportion of students classified in the inconsistent profile of [01] by HO-DINA and DINA. Table 4.3.10 displays those proportions for the six data sets in which HO-DINA and DINA were used to calibrate the data. For LF1223 and PR1223, three consistent profiles [00], [10], and [11] correspond to learning levels 1, 2, and 3, respectively. For LF2334 and PR2334, these profiles represent learning levels-2, -3 and -4. Similarly, they associate with levels-3, -4 and -5 in the last two pairs of CDM data sets (i.e., LF3445 & PR3445). In all data sets, students in profile [01] master higher levels but not the lower ones. Thus, this profile is inconsistent with the theory of learning progressions, if we do not assume that there is an instructional gap between knowledge and skills of lower and higher levels. If the gap exists in the sense that the instruction of higher levels is more recent, students can forget what they had learned of the lower levels, thus could be in profile [01]. Table 4.3.10 clearly showed that for five out of the six data sets, the proportions of students in the inconsistent profile [01] by both CDMs were notably smaller than those from other profiles. The only case where more students were observed in this profile than in one of the three other profiles was the data set for LF item levels 3-4 and 4-5. For this data set, HO-DINA classified six percent of the students as mastering levels 4-5 but not levels 3-4. Whereas, only five percent of all the students were considered to show mastery of levels 3-4 but not levels 4-5 by this model. For the same case, the proportions found using DINA for profiles [01] and [10] were both equal to .01. This evidence aligns with the results described previously regarding item difficulty estimates and attribute locations for this data set. Again, it is noted that there were only six LF level 4-5 items in comparison to nine items measuring the highest levels of PR, and 12 items measuring lowest levels of LF. The small number of items in this LF levels 4-5 group might be a confounding factor reducing the minimum tests' power to detect the true ordering of levels in this case.

Connecting Tables 4.3.9 and 4.3.10, there was a strong relationship between the location distances indicated by HO-DINA and the magnitude of the difference in the proportions of students classified into the inconsistent profile (i.e., [01]) and the other profiles. The more distant the attributes, the more disparate the proportions. In fact, LF2334 had the longest location distance of .73 logits. And, the difference between proportions for the inconsistent profile and the next smallest profile (i.e., [10] in this case) were .25 and .26 using HO-DINA, and DINA, respectively. These differences were also the largest among all the discrepancies of the six data sets. This result can be explained by the nature of the study in that only two attributes were considered at a time. Under this setting, the locations of the attributes will depend largely on the proportion of students in the sample who master each attribute. For attribute-1, the proportion is the sum of proportions of students in profiles [10] and [11]. Similarly, the mastery proportion for attribute-2 can be computed by summing up the statistics of [01] and [11]. If there are fewer students in the inconsistent profile, it is likely that the mastery proportion of attribute-1 will be larger than that of the remaining attribute, which would lead to results indicating the location of the former will be smaller than that of latter.  In what follows, results will be reported to evaluate the level links predicted in the original theory for LF and PR.

### 4.3.3. Claim 2: Co-occurrence of Learning Levels across Progressions

In the theory described in Arieli-Attali et al. (2012), the authors proposed 10 combinations of levels among all the 25 possible level links for LF and PR. Using the notation previously introduced in Table 2.1.2, the postulated level links are (1,1), (1, 2), (2, 2), (2, 3), (3, 3), (3, 4), (4, 4), (4,5), (5, 4) and (5, 5). The plausibility of links (1,1), (1,2), (4,5), (5,4) and (5,5) can be evaluated by using the first and the last data sets. Whereas, the six links in between (1,2) and (4,5) can be examined in two data sets. The following paragraph will describe the investigation in more details.

Tables 4.3.11, 4.3.12, and 4.3.13 report the proportions of students classified into each combination of levels. The word “Yes” in the sixth column was used to indicate a link that was predicted. If the proportions of students classified in this link estimated by MIRT-SS, HO-DINA, or DINA were non-zero, the plausibility for that level link is supported and a check symbol is used in column eight as an indication of that observation. The last column of the tables was used to recommend further considerations for the combinations of levels that were observed in the empirical data through the lens of MIRT-SS and/or the CDMs but not predicted by the theorists. Again, a check mark on a row of a link is used to suggest that follow-up investigations are recommended for the link.

The tables elucidated that all 10 combinations of levels postulated by Arieli-Attali et al. (2012) were observed using either MIRT-SS or the CDMs. Six out of 10 combinations contained more than one percent of the students using all three models. Using the CDM frameworks, students were observed in all 10 combinations, however no students were classified by MIRT-SS in combination (1,2). It is noted that other than the 10 predicted links, MIRT-SS and the CDMs placed non-negligible portions of students in seven other combinations. The most notable links were (2,1) and (3,2). For the combination of level-2 of LF and level-1 of PR, using MIRT-SS, nearly a third of the students were classified in this manner. This result suggested that students in level-2 of LF were more likely to be in level-1 of PR than in any higher levels of this progression. Since the link (2,2) was postulated and supported, the fact that there were more students in combination (2,1) than (2,2) indicated that a large number of students in level-2 of LF were not automatically proficient in knowledge and skills defined by level-2 of PR. If this was the case, the result will have meaningful instructional implications for teachers and students. In short, using MIRT-SS can provide statistical evidence to validate nine out of the 10 level links. This model also suggested the plausibility of three more combinations. HO-DINA and DINA were quite consistent in validating level links. Both models revealed evidence that allowed us to support all the 10 theorized combinations. However, they also suggested that all the 25 combinations were possible. This finding was an illustration of how differently MIRT-SS and CDMs classified students into combination of levels. In the next chapter, the results of this study will be discussed in more detail by connecting them with findings from this dissertation’s two simulation studies as well as published works on the topic of learning progression validation.

As described in Section 3.3.3 of the method chapter, only classification consistency for three pairs of models was collected in the empirical study. The results for this analysis are displayed in Table 4.3.14 toward the end of this chapter. Three themes can be observed from the table. First, the two CDMs seemed to be consistent in classifying students into learning levels. Their consistency rates reached as high as 96% for the data set LFPR1223 and became slightly smaller at 91% for LFPR3445. This finding is expected given the mathematical similarity of HO-

DINA and DINA as shown in Section 2.2.2 of Chapter 2. To calibrate the data using HO-DINA, GDINA estimated the CDM parameters for DINA first. Then, the program used the parameters to estimate attribute locations for the higher-order version of DINA. Second, the consistency between MIRT-SS and HO-DINA or DINA were much lower than the rate within the CDMs. The consistency varied from as low as 39% for MIRT-SS and HO-DINA for LFPR1223 to as high as 65% of MIRT-SS and DINA for LFPR3445. It is noted in the empirical exploration that the consistency rates between MIRT-SS and DINA tended to be slightly higher than those of MIRT-SS and HO-DINA. This finding was different from what was found in Study 1. Nevertheless, the differences were only within 1% to 2% across the three data sets. Last, the consistency rates between MIRT-SS and HO-DINA or DINA in this study were in between the rates found in the simulation investigations. In comparison to the results shown in Tables 4.1.5, 4.2.3, 4.2.5, and 4.2.6, the consistency rates for empirical data seemed to be larger than the rates for the cases of moderate location difference and smaller than these of the extreme location difference. This observation was seen for at least two out of three data sets (i.e., LFPR2334 and LFPR3445). The rates for these data sets were approximately 60% compared to approximately 45% in Study 1, 80% for the extreme difference cases and 52% for the moderate difference cases in Study 2. This result needs to be elaborated and warrants further investigations given the fact that most of the distances of attribute locations in Study 3 were smaller than .5 logits. In the last chapter, this finding will be revisited in more detail.

## 4.4. Chapter Summary

In this chapter, results were reported for all three studies conducted within this dissertation as described in the previous chapters. The results across the studies and conditions considered in each can be summarized using few key observations. First, the five methods used to evaluate the ordering of learning levels seemed to complement each other in the simulation studies. None of the tests to examine learning level order obtained expected true and false positive rates across all simulated conditions. The t-test of item difficulty appeared to be anti-conservative since it resulted in very low false positive rates in most of the cases. Whereas, the order-test, location and minimum tests showed a lack of statistical power to reject null hypothesis in the false scenario. The sensitivity of the methods in detecting the magnitude of level differences was also partially seen in the notable decrease of true positive rates of the t-test between the extreme to moderate difference cases (see Tables 4.2.1 and 4.2.2). In short, the t-test and the remaining methods seemed to perform differently in evaluating the first claim of level ordering.

Second, results of using the models to classify students into level combinations across simulation studies confirmed the mathematical similarity of HO-DINA and DINA and revealed that the consistency between MIRT-SS and the CDMs depended on the magnitude of the differences in difficulty between learning levels. The more distant the levels, the more consistent the model becomes in classifying examinees into level combinations (see Tables 4.2.3 and 4.2.6). Across all true conditions, the cross-model classification consistency between MIRT-SS and the CDMs was far from perfect. This finding illustrates the challenge of using these models to locate students into learning levels and by extension level combinations.

Last, when the models and methods were adopted to analyze empirical data, it was observed that they provided convergent evidence to support almost all aspects of the theory underlying the data. Model-data fit for the MIRT-SS model signified that it fit closely with the empirical study data. Fit information for HO-DINA and DINA was somewhat less promising but deemed acceptable given that it was retrofit to the data. Overall, tests of level ordering based on MIRT-SS and the CDMs supported the theoretical prediction. Using the estimates from fitting the models to the data, students were observed in all 10 theorized combinations (see Tables 4.3.11, 4.3.12, and 4.3.13). In the last chapter, the findings from this empirical study will be discussed at length with respect to the theory underlying LF and PR, the results from Studies 1 and 2, and published works evaluating learning progressions.

## 4.5. Tables and Figures for Chapter 4

**Table 4.1.1. True and False Positive Rates for Tests of Ordered Levels**

| | | | | | Methods | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Scenario | *N* | *I* | *ρ* ($\theta_1$, $\theta_2$) | *Condition* | *MIRT:*<br>*t-tests* | *MIRT:*<br>*Order-test* | *HO-DINA:*<br>*Location test* | *HO-DINA:*<br>*Minimum test* | *DINA:*<br>*Minimum test* |
| True<br>($\mu_{\beta1} < \mu_{\beta2}$) | 500 | 40 | *.6* | 1 | .41 | .95 | .75 | .73 | .85 |
| | | | *.9* | 2 | .51 | .99 | .81 | .80 | .87 |
| | | 60 | *.6* | 3 | .77 | .97 | .86 | .89 | .88 |
| | | | *.9* | 4 | .67 | .96 | .83 | .86 | .90 |
| | 1000 | 40 | *.6* | 5 | .54 | .98 | .88 | .88 | .92 |
| | | | *.9* | 6 | .48 | .97 | .91 | .95 | .95 |
| | | 60 | *.6* | 7 | .71 | 1.00 | .89 | .91 | .93 |
| | | | *.9* | 8 | .80 | 1.00 | .93 | .94 | .98 |
| False<br>($\mu_{\beta1} = \mu_{\beta2}$) | 500 | 40 | *.6* | 9 | .00 | .23 | .27 | .29 | .34 |
| | | | *.9* | 10 | .00 | .29 | .30 | .32 | .40 |
| | | 60 | *.6* | 11 | .00 | .27 | .26 | .29 | .29 |
| | | | *.9* | 12 | .00 | .23 | .27 | .33 | .32 |
| | 1000 | 40 | *.6* | 13 | .00 | .30 | .29 | .41 | .39 |
| | | | *.9* | 14 | .00 | .30 | .32 | .39 | .39 |
| | | 60 | *.6* | 15 | .00 | .26 | .19 | .34 | .20 |
| | | | *.9* | 16 | .01 | .30 | .15 | .29 | .26 |

**Table 4.1.2. True Proportions of Students in Level Combinations**

| Condition | Combinations of Levels | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 11 | 12 | 13 | 21 | 22 | 23 | 31 | 32 | 33 | Total |
| 1 | .18 | .11 | .04 | .09 | .15 | .1 | .04 | .11 | .18 | 1 |
| 2 | .22 | .09 | .01 | .08 | .22 | .08 | .01 | .09 | .22 | 1 |
| 3 | .18 | .11 | .03 | .1 | .16 | .11 | .03 | .1 | .17 | 1 |
| 4 | .23 | .07 | .01 | .09 | .2 | .08 | .01 | .08 | .23 | 1 |
| 5 | .17 | .11 | .03 | .1 | .16 | .1 | .04 | .12 | .17 | 1 |
| 6 | .23 | .09 | .01 | .08 | .19 | .09 | .01 | .09 | .22 | 1 |
| 7 | .18 | .1 | .04 | .1 | .16 | .11 | .03 | .1 | .18 | 1 |
| 8 | .23 | .07 | 0 | .09 | .23 | .09 | .01 | .07 | .22 | 1 |

**Table 4.1.3. Proportions of Students by Level Combination for MIRT**

| Condition | Combinations of Levels | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 11 | 12 | 13 | 21 | 22 | 23 | 31 | 32 | 33 | Total |
| 1 | .18 | .12 | .03 | .09 | .18 | .10 | .03 | .10 | .18 | 1 |
| 2 | .22 | .08 | .00 | .07 | .25 | .07 | .00 | .08 | .21 | 1 |
| 3 | .18 | .11 | .02 | .10 | .19 | .11 | .02 | .10 | .17 | 1 |
| 4 | .23 | .06 | .00 | .08 | .24 | .07 | .00 | .08 | .23 | 1 |
| 5 | .18 | .11 | .02 | .10 | .19 | .09 | .02 | .12 | .17 | 1 |
| 6 | .23 | .08 | .00 | .07 | .24 | .07 | .00 | .08 | .22 | 1 |
| 7 | .19 | .10 | .03 | .10 | .18 | .11 | .02 | .09 | .18 | 1 |
| 8 | .23 | .06 | .00 | .08 | .26 | .07 | .00 | .07 | .22 | 1 |

**Table 4.1.4. Proportions of Students in Each Level Combinations by CDMs**

| Condition | Model | Level Combinations | | | | | | | | | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 11 | 12 | 13 | 21 | 22 | 23 | 31 | 32 | 33 | |
| 1 | HO* | .28 | .04 | .13 | .03 | .01 | .03 | .12 | .04 | .27 | .95 |
| | DINA | .31 | .02 | .15 | .02 | .00 | .02 | .14 | .02 | .30 | .98 |
| 2 | HO | .33 | .04 | .08 | .03 | .01 | .03 | .08 | .04 | .31 | .95 |
| | DINA | .36 | .02 | .10 | .02 | .00 | .02 | .10 | .02 | .34 | .98 |
| 3 | HO | .29 | .03 | .12 | .03 | .01 | .03 | .12 | .03 | .28 | .94 |
| | DINA | .32 | .02 | .14 | .02 | .00 | .02 | .14 | .02 | .30 | .98 |
| 4 | HO | .33 | .03 | .07 | .04 | .01 | .03 | .08 | .03 | .34 | .96 |
| | DINA | .36 | .02 | .08 | .02 | .00 | .02 | .10 | .02 | .36 | .98 |
| 5 | HO | .28 | .04 | .12 | .04 | .01 | .04 | .12 | .04 | .27 | .96 |
| | DINA | .30 | .02 | .15 | .02 | .00 | .02 | .14 | .02 | .30 | .97 |
| 6 | HO | .32 | .04 | .08 | .03 | .01 | .03 | .08 | .04 | .32 | .95 |
| | DINA | .35 | .02 | .10 | .02 | .00 | .02 | .10 | .02 | .35 | .98 |
| 7 | HO | .28 | .03 | .12 | .03 | .01 | .03 | .12 | .03 | .28 | .93 |
| | DINA | .31 | .02 | .14 | .02 | .00 | .02 | .14 | .02 | .31 | .98 |
| 8 | HO | .33 | .03 | .07 | .04 | .01 | .03 | .07 | .03 | .34 | .95 |
| | DINA | .36 | .02 | .09 | .02 | .00 | .02 | .09 | .02 | .36 | .98 |

*) HO-DINA model (HO is used in some following tables due to limited space)

**Table 4.1.5. Proportions of Students in Inconsistent Combinations**

| Condition | Model | Inconsistent Combinations | | | | | | | Total |
|---|---|---|---|---|---|---|---|---|---|
| | | 1I | I1 | 2I | I2 | 3I | I3 | II | |
| 1 | HO | .01 | .01 | .00 | .00 | .01 | .01 | .00 | .04 |
| | DINA | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 |
| 2 | HO | .01 | .01 | .00 | .00 | .01 | .01 | .00 | .04 |
| | DINA | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 |
| 3 | HO | .01 | .01 | .00 | .00 | .01 | .01 | .00 | .04 |
| | DINA | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 |
| 4 | HO | .01 | .01 | .00 | .00 | .01 | .01 | .00 | .04 |
| | DINA | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 |
| 5 | HO | .01 | .01 | .00 | .00 | .01 | .01 | .00 | .04 |
| | DINA | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 |
| 6 | HO | .01 | .01 | .00 | .00 | .01 | .01 | .00 | .04 |
| | DINA | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 |
| 7 | HO | .01 | .01 | .00 | .00 | .01 | .01 | .00 | .04 |
| | DINA | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 |
| 8 | HO | .01 | .01 | .00 | .00 | .01 | .01 | .00 | .04 |
| | DINA | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 |

**Table 4.1.6. Classification Accuracy for True Conditions**

| Condition | Accuracy Rate | | | | | |
|---|---|---|---|---|---|---|
| | MIRT Ordered Cases | MIRT Significant Cases | HO-DINA Ordered Cases | HO-DINA Significant Cases | DINA Ordered Cases | DINA Significant Cases |
| 1 | .60 | .61 | .40 | .36 | .39 | .34 |
| 2 | .65 | .65 | .42 | .37 | .42 | .36 |
| 3 | .64 | .64 | .41 | .40 | .40 | .38 |
| 4 | .68 | .67 | .46 | .44 | .45 | .44 |
| 5 | .61 | .61 | .39 | .37 | .38 | .36 |
| 6 | .65 | .65 | .44 | .39 | .44 | .38 |
| 7 | .65 | .65 | .43 | .41 | .42 | .39 |
| 8 | .69 | .69 | .46 | .44 | .45 | .43 |

**Table 4.1.7. Cross-model Classification Consistency for True Conditions**

| Condition | Consistency Rate | | | |
|---|---|---|---|---|
| | HO-DINA vs. MIRT: Ordered Cases | HO-DINA vs. MIRT: Significant Cases | DINA vs. MIRT: Ordered Cases | DINA vs. MIRT: Significant Cases |
| 1 | .45 | .39 | .43 | .37 |
| 2 | .46 | .40 | .45 | .39 |
| 3 | .45 | .42 | .42 | .40 |
| 4 | .49 | .47 | .48 | .46 |
| 5 | .43 | .40 | .41 | .38 |
| 6 | .47 | .41 | .46 | .40 |
| 7 | .47 | .44 | .45 | .41 |
| 8 | .49 | .47 | .48 | .46 |

**Table 4.2.1. True and False Positive Rates for Extreme Difference Cases**

| Condition | MIRT:<br>t-test | MIRT:<br>order- test | HO-DINA:<br>Location<br>Test | HO-DINA:<br>minimum<br>test | DINA:<br>minimum<br>test |
|---|---|---|---|---|---|
| 1 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 2 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 3 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 4 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 5 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 6 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 7 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 8 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 9 | .01 | .29 | .22 | .29 | .31 |
| 10 | .02 | .27 | .25 | .35 | .34 |
| 11 | .01 | .23 | .23 | .28 | .28 |
| 12 | .00 | .24 | .22 | .29 | .29 |
| 13 | .00 | .2 | .26 | .3 | .32 |
| 14 | .00 | .21 | .24 | .32 | .33 |
| 15 | .00 | .26 | .19 | .31 | .31 |
| 16 | .02 | .28 | .25 | .35 | .34 |

**Table 4.2.2. True and False Positive Rates for Moderate Difference Cases**

| Condition | MIRT: t-test | MIRT: order-test | HO-DINA: Location Test | HO-DINA: minimum test | DINA: minimum test |
|---|---|---|---|---|---|
| 1 | .55 | .91 | 1.00 | 1.00 | 1.00 |
| 2 | .54 | .93 | 1.00 | 1.00 | 1.00 |
| 3 | .67 | .96 | 1.00 | 1.00 | 1.00 |
| 4 | .63 | .96 | 1.00 | 1.00 | 1.00 |
| 5 | .51 | .89 | 1.00 | 1.00 | 1.00 |
| 6 | .49 | .91 | 1.00 | 1.00 | 1.00 |
| 7 | .70 | .97 | 1.00 | 1.00 | 1.00 |
| 8 | .78 | .95 | 1.00 | 1.00 | 1.00 |
| 9 | .01 | .23 | .22 | .31 | .31 |
| 10 | .02 | .34 | .26 | .35 | .35 |
| 11 | .01 | .26 | .28 | .35 | .34 |
| 12 | .03 | .28 | .25 | .35 | .36 |
| 13 | .01 | .22 | .16 | .29 | .26 |
| 14 | .01 | .25 | .24 | .29 | .29 |
| 15 | .01 | .22 | .24 | .35 | .36 |
| 16 | .01 | .2 | .23 | .36 | .36 |

**Table 4.2.3. Classification Accuracy and Consistency for Extreme Difference Cases**

| Condition | Classification Accuracy Rate | | | Cross-model Consistency Rate | |
|---|---|---|---|---|---|
| | HO vs. True | MIRT vs. True | DINA vs. True | MIRT vs. HO | DINA vs. HO |
| 1 | .97 | .75 | .97 | .75 | 1.00 |
| 2 | .97 | .76 | .97 | .76 | 1.00 |
| 3 | .99 | .81 | .99 | .81 | 1.00 |
| 4 | .99 | .83 | .99 | .83 | 1.00 |
| 5 | .97 | .75 | .97 | .76 | 1.00 |
| 6 | .97 | .75 | .97 | .75 | 1.00 |
| 7 | .99 | .83 | .99 | .83 | 1.00 |
| 8 | .99 | .83 | .99 | .83 | 1.00 |

**Table 4.2.4. Proportions of Students in Level Links (Extreme Difference Cases)**

| Condition | Model | Combinations of Levels | | | | | | | | | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 11 | 12 | 13 | 21 | 22 | 23 | 31 | 32 | 33 | |
| 1 | TRUE | .08 | .12 | .02 | .12 | .29 | .11 | .02 | .11 | .08 | .95 |
| | HO | .08 | .12 | .03 | .12 | .29 | .11 | .02 | .11 | .08 | .96 |
| | MIRT | .11 | .13 | .02 | .14 | .32 | .10 | .03 | .10 | .05 | 1.00 |
| | DINA | .08 | .12 | .03 | .12 | .29 | .11 | .02 | .11 | .08 | .96 |
| 2 | TRUE | .10 | .11 | .01 | .10 | .31 | .11 | .01 | .11 | .10 | .96 |
| | HO | .10 | .11 | .01 | .11 | .31 | .11 | .02 | .11 | .10 | .98 |
| | MIRT | .12 | .13 | .02 | .12 | .33 | .10 | .01 | .10 | .07 | 1.00 |
| | DINA | .10 | .11 | .01 | .11 | .31 | .11 | .02 | .11 | .10 | .98 |
| 3 | TRUE | .08 | .11 | .02 | .11 | .30 | .11 | .02 | .11 | .08 | .94 |
| | HO | .08 | .11 | .02 | .12 | .30 | .11 | .02 | .11 | .08 | .95 |
| | MIRT | .10 | .13 | .03 | .13 | .31 | .11 | .02 | .11 | .06 | 1.00 |
| | DINA | .08 | .11 | .02 | .12 | .30 | .11 | .02 | .11 | .08 | .95 |
| 4 | TRUE | .10 | .11 | .01 | .11 | .31 | .11 | .01 | .10 | .10 | .96 |
| | HO | .10 | .11 | .01 | .11 | .31 | .11 | .01 | .10 | .10 | .96 |
| | MIRT | .12 | .12 | .01 | .12 | .33 | .10 | .02 | .10 | .08 | 1.00 |
| | DINA | .10 | .11 | .01 | .11 | .31 | .11 | .01 | .10 | .10 | .96 |
| 5 | TRUE | .08 | .11 | .02 | .11 | .3 | .11 | .02 | .12 | .08 | .95 |
| | HO | .08 | .11 | .02 | .11 | .3 | .11 | .03 | .12 | .08 | .96 |
| | MIRT | .11 | .14 | .02 | .13 | .32 | .10 | .02 | .10 | .06 | 1.00 |
| | DINA | .08 | .11 | .02 | .11 | .3 | .11 | .03 | .12 | .08 | .96 |
| 6 | TRUE | .10 | .11 | .01 | .11 | .31 | .11 | .01 | .11 | .10 | .97 |
| | HO | .10 | .11 | .01 | .11 | .31 | .11 | .01 | .11 | .10 | .97 |
| | MIRT | .13 | .12 | .01 | .13 | .32 | .10 | .02 | .10 | .07 | 1.00 |
| | DINA | .10 | .11 | .01 | .11 | .31 | .11 | .01 | .11 | .10 | .97 |
| 7 | TRUE | .08 | .11 | .02 | .11 | .3 | .11 | .02 | .12 | .08 | .95 |
| | HO | .08 | .12 | .02 | .11 | .3 | .11 | .02 | .12 | .08 | .96 |
| | MIRT | .10 | .13 | .02 | .13 | .31 | .11 | .03 | .11 | .06 | 1.00 |
| | DINA | .08 | .12 | .02 | .11 | .3 | .11 | .02 | .12 | .08 | .96 |
| 8 | TRUE | .10 | .11 | .01 | .11 | .31 | .11 | .01 | .11 | .10 | .97 |
| | HO | .10 | .11 | .01 | .11 | .31 | .11 | .01 | .10 | .10 | .96 |
| | MIRT | .12 | .12 | .01 | .12 | .33 | .11 | .01 | .10 | .08 | 1.00 |
| | DINA | .10 | .11 | .01 | .11 | .31 | .11 | .01 | .10 | .10 | .96 |

**Table 4.2.5. Proportions of Students in Inconsistent Links (Extreme Difference Cases)**

| Condition | Model | Inconsistent Combinations | | | | | | | Total |
|---|---|---|---|---|---|---|---|---|---|
| | | 1I* | I1 | 2I | I2 | 3I | I3 | II | |
| 1 | TRUE | .00 | .00 | .01 | .01 | .00 | .00 | .00 | .02 |
| | HO | .00 | .00 | .01 | .01 | .00 | .00 | .00 | .02 |
| | DINA | .00 | .00 | .01 | .01 | .00 | .00 | .00 | .02 |
| 2 | TRUE | .00 | .00 | .01 | .01 | .00 | .00 | .00 | .02 |
| | HO | .00 | .00 | .01 | .01 | .00 | .00 | .00 | .02 |
| | DINA | .00 | .00 | .01 | .01 | .00 | .00 | .00 | .02 |
| 3 | TRUE | .00 | .00 | .01 | .01 | .00 | .00 | .00 | .02 |
| | HO | .00 | .00 | .01 | .01 | .00 | .00 | .00 | .02 |
| | DINA | .00 | .00 | .01 | .01 | .00 | .00 | .00 | .02 |
| 4 | TRUE | .00 | .00 | .01 | .01 | .00 | .00 | .00 | .02 |
| | HO | .00 | .00 | .01 | .01 | .00 | .00 | .00 | .02 |
| | DINA | .00 | .00 | .01 | .01 | .00 | .00 | .00 | .02 |
| 5 | TRUE | .00 | .00 | .01 | .01 | .00 | .00 | .00 | .02 |
| | HO | .00 | .00 | .01 | .01 | .00 | .00 | .00 | .02 |
| | DINA | .00 | .00 | .01 | .01 | .00 | .00 | .00 | .02 |
| 6 | TRUE | .00 | .00 | .01 | .01 | .00 | .00 | .00 | .02 |
| | HO | .00 | .00 | .01 | .01 | .00 | .00 | .00 | .02 |
| | DINA | .00 | .00 | .01 | .01 | .00 | .00 | .00 | .02 |
| 7 | TRUE | .00 | .00 | .01 | .01 | .00 | .00 | .00 | .02 |
| | HO | .00 | .00 | .01 | .01 | .00 | .00 | .00 | .02 |
| | DINA | .00 | .00 | .01 | .01 | .00 | .00 | .00 | .02 |
| 8 | TRUE | .00 | .00 | .01 | .01 | .00 | .00 | .00 | .02 |
| | HO | .00 | .00 | .01 | .01 | .00 | .00 | .00 | .02 |
| | DINA | .00 | .00 | .01 | .01 | .00 | .00 | .00 | .02 |

*) I: inconsistent profile [01]

**Table 4.2.6. Accuracy and Consistency for Moderate Difference Cases (Significant Cases)**

| Condition | Classification Accuracy | | | Cross-model Consistency | |
|---|---|---|---|---|---|
| | HO vs. True | MIRT vs. True | DINA vs. True | MIRT vs. HO | DINA vs. HO |
| 1 | .97 | .51 | .97 | .51 | 1.00 |
| 2 | .97 | .51 | .97 | .51 | 1.00 |
| 3 | .99 | .53 | .99 | .53 | 1.00 |
| 4 | .99 | .52 | .99 | .52 | 1.00 |
| 5 | .97 | .50 | .97 | .51 | 1.00 |
| 6 | .97 | .52 | .97 | .52 | 1.00 |
| 7 | .99 | .52 | .99 | .53 | 1.00 |
| 8 | .99 | .53 | .99 | .53 | 1.00 |

**Table 4.2.7. Accuracy and Consistency for Moderate Difference Cases (Ordered Cases)**

| Conditions | Classification Accuracy | | | Cross-model Consistency | |
|---|---|---|---|---|---|
| | HO vs. True | MIRT vs. True | DINA vs. True | MIRT vs. HO | DINA vs. HO |
| 1 | .97 | .49 | .97 | .49 | 1.00 |
| 2 | .97 | .50 | .97 | .50 | 1.00 |
| 3 | .99 | .52 | .99 | .52 | 1.00 |
| 4 | .99 | .51 | .99 | .51 | 1.00 |
| 5 | .97 | .49 | .97 | .49 | 1.00 |
| 6 | .97 | .50 | .97 | .50 | 1.00 |
| 7 | .99 | .51 | .99 | .51 | 1.00 |
| 8 | .99 | .52 | .99 | .52 | 1.00 |

**Table 4.2.8. Proportions of Students in Level Links (Moderate Difference Cases)**

| Condition | Model | Level Combinations | | | | | | | | | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 11 | 12 | 13 | 21 | 22 | 23 | 31 | 32 | 33 | |
| 1 | TRUE | .16 | .08 | .06 | .08 | .06 | .08 | .06 | .08 | .15 | .81 |
| | HO | .16 | .08 | .06 | .08 | .06 | .08 | .06 | .08 | .15 | .81 |
| | MIRT | .34 | .06 | .13 | .07 | .02 | .05 | .13 | .05 | .16 | 1.00 |
| | DINA | .16 | .08 | .06 | .08 | .06 | .08 | .06 | .08 | .15 | .81 |
| 2 | TRUE | .18 | .07 | .04 | .08 | .07 | .07 | .04 | .07 | .18 | .80 |
| | HO | .18 | .07 | .04 | .08 | .07 | .07 | .04 | .07 | .18 | .80 |
| | MIRT | .36 | .06 | .10 | .07 | .02 | .05 | .10 | .05 | .18 | 1.00 |
| | DINA | .18 | .07 | .04 | .08 | .07 | .07 | .04 | .07 | .18 | .80 |
| 3 | TRUE | .15 | .08 | .06 | .08 | .06 | .08 | .06 | .08 | .15 | .80 |
| | HO | .15 | .08 | .06 | .08 | .06 | .08 | .06 | .08 | .15 | .80 |
| | MIRT | .33 | .06 | .14 | .06 | .01 | .04 | .14 | .04 | .17 | 1.00 |
| | DINA | .15 | .08 | .06 | .08 | .06 | .08 | .06 | .08 | .15 | .80 |
| 4 | TRUE | .19 | .07 | .04 | .07 | .07 | .07 | .04 | .07 | .18 | .80 |
| | HO | .18 | .07 | .04 | .07 | .07 | .07 | .04 | .07 | .18 | .79 |
| | MIRT | .37 | .06 | .11 | .05 | .01 | .04 | .11 | .05 | .19 | 1.00 |
| | DINA | .18 | .07 | .04 | .07 | .07 | .07 | .04 | .07 | .18 | .79 |
| 5 | TRUE | .16 | .08 | .06 | .08 | .06 | .08 | .06 | .08 | .15 | .81 |
| | HO | .16 | .08 | .06 | .08 | .06 | .08 | .06 | .08 | .15 | .81 |
| | MIRT | .34 | .06 | .13 | .07 | .02 | .05 | .13 | .05 | .16 | 1.00 |
| | DINA | .16 | .08 | .06 | .08 | .06 | .08 | .06 | .08 | .15 | .81 |
| 6 | TRUE | .18 | .07 | .04 | .07 | .07 | .07 | .04 | .07 | .18 | .79 |
| | HO | .18 | .08 | .04 | .08 | .07 | .07 | .04 | .07 | .18 | .81 |
| | MIRT | .36 | .06 | .11 | .06 | .02 | .05 | .11 | .05 | .19 | 1.00 |
| | DINA | .18 | .08 | .04 | .08 | .07 | .07 | .04 | .07 | .18 | .81 |
| 7 | TRUE | .16 | .08 | .06 | .08 | .06 | .08 | .06 | .08 | .15 | .81 |
| | HO | .16 | .08 | .06 | .08 | .06 | .08 | .06 | .08 | .15 | .81 |
| | MIRT | .34 | .06 | .14 | .06 | .02 | .04 | .14 | .04 | .17 | 1.00 |
| | DINA | .16 | .08 | .06 | .08 | .06 | .08 | .06 | .08 | .15 | .81 |
| 8 | TRUE | .18 | .07 | .04 | .07 | .07 | .08 | .04 | .07 | .18 | .80 |
| | HO | .18 | .07 | .04 | .08 | .07 | .08 | .04 | .07 | .18 | .81 |
| | MIRT | .36 | .06 | .11 | .06 | .02 | .04 | .10 | .05 | .20 | 1.00 |
| | DINA | .18 | .07 | .04 | .08 | .07 | .08 | .04 | .07 | .18 | .81 |

**Table 4.2.8. Proportions of Students in Inconsistent Links (Moderate Difference Cases)**

| Condition | Model | Inconsistent Combinations | | | | | | | Total |
|---|---|---|---|---|---|---|---|---|---|
| | | 1I* | I1 | 2I | I2 | 3I | I3 | II | |
| 1 | TRUE | .03 | .03 | .03 | .03 | .03 | .03 | .01 | .19 |
| | HO | .03 | .03 | .03 | .03 | .03 | .03 | .01 | .19 |
| | DINA | .03 | .03 | .03 | .03 | .03 | .03 | .01 | .19 |
| 2 | TRUE | .03 | .03 | .03 | .03 | .03 | .03 | .01 | .19 |
| | HO | .03 | .03 | .03 | .03 | .03 | .03 | .01 | .19 |
| | DINA | .03 | .03 | .03 | .03 | .03 | .03 | .01 | .19 |
| 3 | TRUE | .03 | .03 | .03 | .03 | .03 | .03 | .01 | .19 |
| | HO | .03 | .03 | .03 | .03 | .03 | .03 | .01 | .19 |
| | DINA | .03 | .03 | .03 | .03 | .03 | .03 | .01 | .19 |
| 4 | TRUE | .03 | .03 | .03 | .03 | .03 | .03 | .01 | .19 |
| | HO | .03 | .03 | .03 | .03 | .03 | .03 | .01 | .19 |
| | DINA | .03 | .03 | .03 | .03 | .03 | .03 | .01 | .19 |
| 5 | TRUE | .03 | .03 | .03 | .03 | .03 | .03 | .01 | .19 |
| | HO | .03 | .03 | .03 | .03 | .03 | .03 | .01 | .19 |
| | DINA | .03 | .03 | .03 | .03 | .03 | .03 | .01 | .19 |
| 6 | TRUE | .03 | .03 | .03 | .03 | .03 | .03 | .01 | .19 |
| | HO | .03 | .03 | .03 | .03 | .03 | .03 | .01 | .19 |
| | DINA | .03 | .03 | .03 | .03 | .03 | .03 | .01 | .19 |
| 7 | TRUE | .03 | .03 | .03 | .03 | .03 | .03 | .01 | .19 |
| | HO | .03 | .03 | .03 | .03 | .03 | .03 | .01 | .19 |
| | DINA | .03 | .03 | .03 | .03 | .03 | .03 | .01 | .19 |
| 8 | TRUE | .03 | .03 | .03 | .03 | .03 | .03 | .01 | .19 |
| | HO | .03 | .03 | .03 | .03 | .03 | .03 | .01 | .19 |
| | DINA | .03 | .03 | .03 | .03 | .03 | .03 | .01 | .19 |

*) I: inconsistent profile [01]

**Table 4.3.1. Limited-Information Fit Statistics and Indexes for MIRT-SS**

| Data Set | $M_2$ | Df | Prob | $\widehat{F_0}$ | RMSEA | TLI | Conclusion |
|---|---|---|---|---|---|---|---|
| LFPR1223 | 4461.82 | 3259 | .0001 | 7.80 | .03 | .85 | Close Fit |
| LFPR2334 | 4579.79 | 2976 | .0001 | 7.52 | .03 | .92 | Close Fit |
| LFPR3445 | 2536.32 | 1581 | .0001 | 3.18 | .03 | .90 | Close Fit |

**Table 4.3.2. Summary Chen-Thissen LD $X^2$ Fit Statistics for MIRT-SS**

| Data set | Percent within [-3,3] | Conclusion |
|---|---|---|
| LFPR1223 | 92.5 | Good Fit |
| LFPR2334 | 97.2 | Good Fit |
| LFPR3445 | 91.2 | Good Fit |

**Table 4.3.3. Relative Fit Statistics for Fitting CDMs to Empirical Data**

| Data set | HO-DINA | | | DINA | | | Selected Model |
|---|---|---|---|---|---|---|---|
| | AIC | BIC | N. par | AIC | BIC | N. par | |
| LF1223 | 10452.04 | 1106.91 | 140 | 10445.72 | 1105.25 | 139 | DINA |
| LF2334 | 1303.19 | 13815.50 | 178 | 13028.09 | 13808.98 | 177 | DINA |
| LF3445 | 9564.42 | 9938.89 | 80 | 9556.51 | 9926.30 | 79 | DINA |
| PR1223 | 9115.18 | 9532.70 | 96 | 9105.50 | 9518.66 | 95 | DINA |
| PR2334 | 11323.17 | 11984.94 | 150 | 11308.46 | 11965.82 | 149 | DINA |
| PR3445 | 10527.34 | 10976.70 | 96 | 10516.04 | 1096.72 | 95 | DINA |

**Table 4.3.4. Absolute Fit Statistics for Fitting HO-DINA to Empirical Data**

| Data set | Proportion Correct | | | Transformed Correlation | | | Log odds ratio | | |
|---|---|---|---|---|---|---|---|---|---|
| | Max z | P | Adj. p | Max z | p | Adj. p | Max z | p | Adj. p |
| LF1223 | 2.51 | .01 | .83 | 41.32 | .00 | .00 | 3.31 | .00 | .00 |
| LF2334 | 4.79 | .00 | .00 | 16.16 | .00 | .00 | 11.93 | .00 | .00 |
| LF3445 | 2.61 | .01 | .34 | 1.67 | .00 | .00 | 1.20 | .00 | .00 |
| PR1223 | 1.94 | .05 | 1 | 39.83 | .00 | .00 | 26.42 | .00 | .00 |
| PR2334 | 1.13 | .26 | 1 | 37.33 | .00 | .00 | 24.34 | .00 | .00 |
| PR3445 | 3.10 | .00 | .09 | 33.41 | .00 | .00 | 39.35 | .00 | .00 |

**Table 4.3.5. Absolute Fit Statistics for Fitting DINA to Empirical Data**

| Data set | Proportion Correct | | | Transformed Correlation | | | Log odds ratio | | |
|---|---|---|---|---|---|---|---|---|---|
| | Max z | P | Adj. p | Max z | P | Adj. p | Max z | p | Adj. p |
| LF1223 | 2.71 | .00 | .46 | 41.38 | .00 | .00 | 3.77 | .00 | .00 |
| LF2334 | 4.68 | .00 | .00 | 16.13 | .00 | .00 | 11.91 | .00 | .00 |
| LF3445 | 2.76 | .00 | .22 | 1.87 | .00 | .00 | 1.42 | .00 | .00 |
| PR1223 | 2.06 | .04 | 1 | 39.88 | .00 | .00 | 26.44 | .00 | .00 |
| PR2334 | 1.63 | .10 | 1 | 35.40 | .00 | .00 | 24.39 | .00 | .00 |
| PR3445 | 3.05 | .00 | .10 | 33.46 | .00 | .00 | 39.36 | .00 | .00 |

**Table 4.3.6. Summary of Item Parameter Estimates by Fitting CDMs to Empirical Data**

| Data set | HO-DINA | | | | DINA | | | |
|---|---|---|---|---|---|---|---|---|
| | Guessing (g) | | Slipping (s) | | Guessing (g) | | Slipping (s) | |
| | Mean/SD | % < .4 | Mean/SD | % < .4 | Mean/SD | % < .4 | Mean/SD | % < .4 |
| LF1223 | .26/.26 | 72 | .40/.30 | 56 | .26/.26 | 72 | .40/.30 | 56 |
| LF2334 | .24/.23 | 78 | .36/.24 | 54 | .24/.23 | 78 | .36/.24 | 54 |
| LF3445 | .23/.15 | 84 | .29/.21 | 71 | .23/.15 | 84 | .29/.21 | 71 |
| PR1223 | .25/.21 | 76 | .35/.24 | 57 | .25/.21 | 76 | .34/.24 | 61 |
| PR2334 | .20/.17 | 85 | .40/.23 | 51 | .20/.19 | 86 | .40/.23 | 51 |
| PR3445 | .20/.19 | 80 | .36/.23 | 61 | .20/.19 | 80 | .36/.23 | 59 |

**Table 4.3.7. Summary of Discrimination Indexes of CDMs for Empirical Data**

| Data set | HO-DINA | | | DINA | | |
|---|---|---|---|---|---|---|
| | Discrimination ($\delta$ = 1- s – g) | | | Discrimination ($\delta$ = 1- s – g) | | |
| | Mean/SD | % > .2 | range | Mean/SD | % > .2 | range |
| LF1223 | .34/.17 | 76 | [.02, .71] | .34/.17 | 76 | [.02, .71] |
| LF2334 | .40/.17 | 86 | [.10, .92] | .39/.17 | 86 | [.10, .91] |
| LF3445 | .49/.20 | 89 | [.10, .81] | .48/.20 | 87 | [.10, .81] |
| PR1223 | .40/.17 | .87 | [.04, .81] | .40/.18 | 87 | [.04, .80] |
| PR2334 | .41/.16 | 92 | [.05, .72] | .40/.16 | 92 | [.04, .73] |
| PR3445 | .44/.20 | 89 | [-.01, .77] | .44/.20 | 89 | [-.02, .78] |

**Table 4.3.8. Results of Two-sample T-tests Comparing Item Difficulties**

| Pair | Number of items | Mean (SD) | t | df | p | Cohen-d |
|---|---|---|---|---|---|---|
| LF12 vs. LF23 | 12 vs. 56 | -1.17 (1.74) vs. 0.76 (1.61) | -3.54 | 15.3 | .00* | 1.19 |
| LF23 vs. LF34 | 58 vs. 29 | 0.10 (1.18) vs. 1.52 (1.00) | -5.84 | 64.9 | .00* | 1.25 |
| LF34 vs. LF45 | 32 vs. 6 | 0.39 (.91) vs. 1.01 (.82) | -1.69 | 7.4 | .07 | *.70* |
| PR12 vs. PR23 | 14 vs. 32 | -0.03 (1.3) vs. 0.79 (1.20) | -2.00 | 23 | .03* | .66 |
| PR23 vs. PR34 | 32 vs. 41 | 0.29 (.75) vs. 0.81 (1.01) | -2.50 | 7.8 | .01* | .57 |
| PR34 vs. PR45 | 40 vs. 9 | 0.09 (2.05) vs. 1.00 (1.11) | -1.82 | 23.1 | .04* | .48 |

*) significant at alpha level of .05

**Table 4.3.9. Attribute Locations for Six Data Sets**

| Data sets | Locations |  |  | Support the theory |
|---|---|---|---|---|
|  | Attribute 1 | Attribute 2 | Distance |  |
| LF1223 | -.35 | .11 | .46 | ✓ |
| LF2334 | .04 | .77 | .73 | ✓ |
| LF3445 | .37 | .32 | -.05 | ✗ |
| PR1223 | .24 | .29 | .05 | ✓ |
| PR2334 | .13 | .29 | .16 | ✓ |
| PR3445 | .20 | .59 | .39 | ✓ |

**Table 4.3.10. Proportions of Students in the Learning Profiles by CDMs**

| Data sets | HO-DINA |  |  |  | DINA |  |  |  | Support the theory |
|---|---|---|---|---|---|---|---|---|---|
|  | [00] | [10] | [01] | [11] | [00] | [10] | [01] | [11] |  |
| LF1223 | .36 | .09 | .01 | .54 | .37 | .09 | .01 | .54 | ✓ |
| LF2334 | .51 | .26 | .01 | .22 | .51 | .26 | 0 | .22 | ✓ |
| LF3445 | .59 | *.05* | *.06* | .31 | .63 | *.01* | *.01* | .35 | ✗ |
| PR1223 | .56 | .05 | .04 | .35 | .57 | .05 | .02 | .37 | ✓ |
| PR2334 | .52 | .09 | .03 | .36 | .53 | .08 | .01 | .39 | ✓ |
| PR3445 | .56 | .15 | .01 | .28 | .57 | .13 | .01 | .29 | ✓ |

**Table 4.3.11. Proportions of Students in Level Combinations (LFPR1223)**

| Levels | | | Percentages of students | | | Postulated | Supported | Further Consideration |
|---|---|---|---|---|---|---|---|---|
| Links | LF | PR | MIRT | HO-DINA | DINA | | | |
| (1,1) | 1 | 1 | 8.74 | 31.29 | 31.82 | Yes | ✓ | |
| (1,2) | 1 | 2 | 0.00 | 1.05 | 1.05 | Yes | ✓ | |
| (1,3) | 1 | 3 | 0.00 | 2.62 | 3.15 | | | |
| (2,1) | 2 | 1 | 35.90 | 5.77 | 5.94 | | | ✓ |
| (2,2) | 2 | 2 | 28.90 | .52 | 0.35 | Yes | ✓ | |
| (2,3) | 2 | 3 | 2.10 | 2.27 | 1.92 | Yes | ✓ | |
| (3,1) | 3 | 1 | 0.70 | 18.36 | 18.53 | | | ✓ |
| (3,2) | 3 | 2 | 11.54 | 3.32 | 3.32 | | | ✓ |
| (3,3) | 3 | 3 | 25.35 | 29.90 | 31.64 | Yes | ✓ | |

**Table 4.3.12. Proportions of Students in Level Combinations (LFPR2334)**

| Level | | | Percentages of students | | | Postulated | Supported | Further consideration |
|---|---|---|---|---|---|---|---|---|
| Links | LF | PR | MIRT | HO-DINA | DINA | | | |
| (2,2) | 2 | 2 | 44.16 | 39.24 | 39.9 | Yes | ✓ | |
| (2,3) | 2 | 3 | 2.30 | 3.45 | 2.79 | Yes | ✓ | |
| (2,4) | 2 | 4 | 0.33 | 6.90 | 8.37 | | | ✓ |
| (3,2) | 3 | 2 | 16.26 | 8.05 | 9.03 | | | ✓ |
| (3,3) | 3 | 3 | 16.42 | 4.11 | 3.45 | Yes | ✓ | |
| (3,4) | 3 | 4 | 12.48 | 11.99 | 13.3 | Yes | ✓ | |
| (4,2) | 4 | 2 | 0.00 | 4.11 | 3.94 | | | |
| (4,3) | 4 | 3 | 0.99 | 1.64 | 1.48 | | | |
| (4,4) | 4 | 4 | 6.08 | 16.09 | 16.58 | Yes | ✓ | |

**Table 4.3.13. Proportions of Students in Level Combinations (LFPR3445)**

| Level | | | Percentages of students | | | Postulated | Supported | Further consideration |
|---|---|---|---|---|---|---|---|---|
| Links | LF | PR | MIRT | HO-DINA | DINA | | | |
| (3,3) | 3 | 3 | 57.97 | 44.67 | 47.68 | Yes | ✓ | |
| (3,4) | 3 | 4 | 3.51 | 8.03 | 7.65 | Yes | ✓ | |
| (3,5) | 3 | 5 | 1.13 | 5.40 | 6.9 | | | ✓ |
| (4,3) | 4 | 3 | 8.03 | 2.13 | 0.50 | | | ✓ |
| (4,4) | 4 | 4 | 6.65 | 0.50 | 0.13 | Yes | ✓ | |
| (4,5) | 4 | 5 | 5.77 | 1.88 | 0.38 | Yes | ✓ | |
| (5,3) | 5 | 3 | 0.25 | 6.40 | 8.53 | | | |
| (5,4) | 5 | 4 | 2.89 | 5.27 | 5.27 | Yes | ✓ | |
| (5,5) | 5 | 5 | 13.80 | 19.07 | 21.08 | Yes | ✓ | |

**Table 4.3.14. Decision Consistency Between Pairs of Models**

| Data Set | MIRT vs. HO-DINA | MIRT vs. DINA | HO-DINA vs. DINA |
|---|---|---|---|
| LFPR1223 | .39 | .40 | .96 |
| LFPR2334 | .57 | .59 | .94 |
| LFPR3445 | .64 | .65 | .91 |

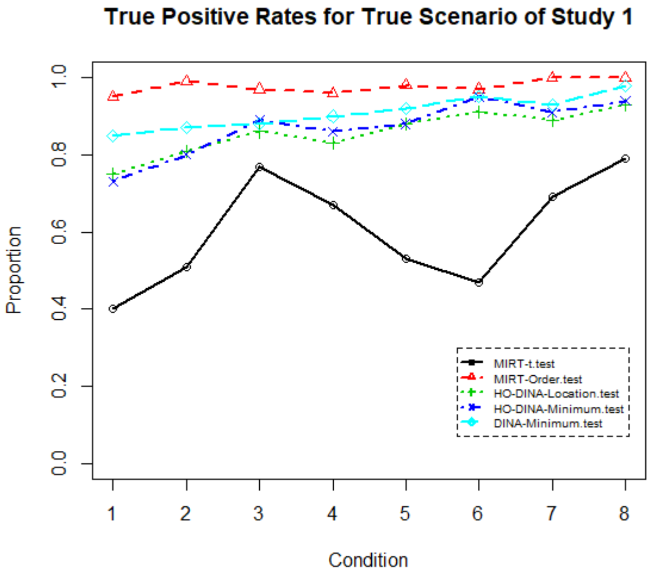


**Figure 4.1.1. True Positive Rates by Five Methods for True Scenario**

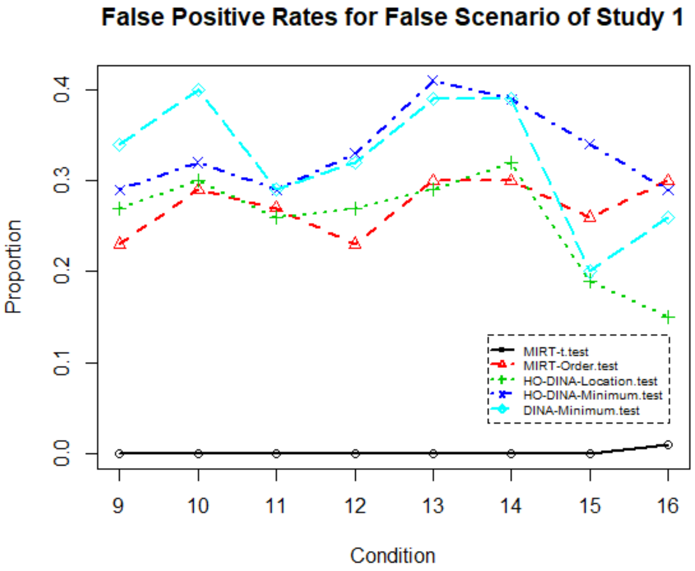
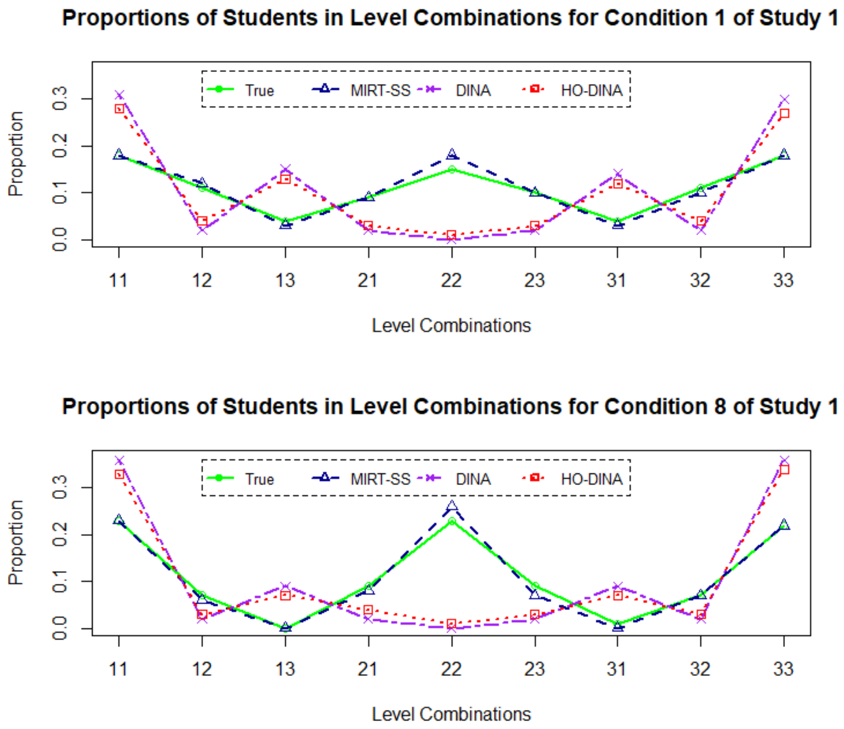
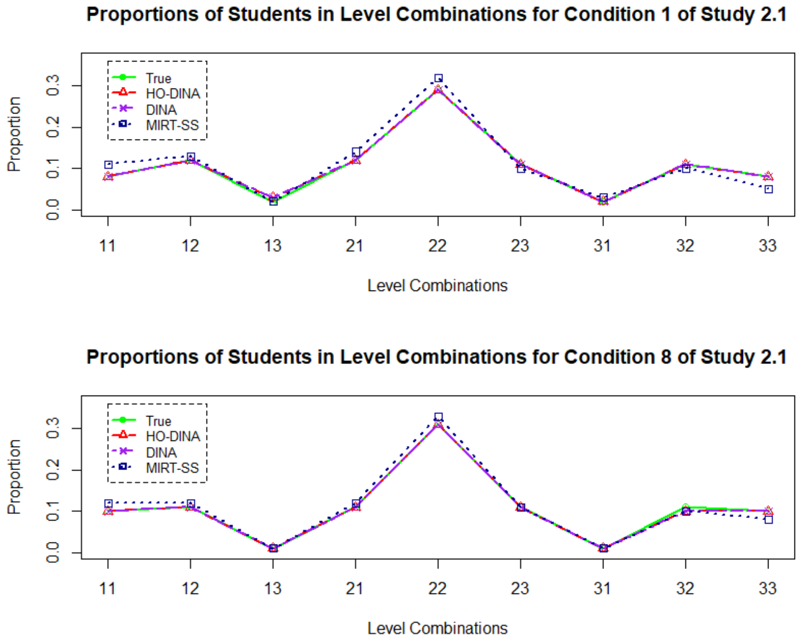


**Figure 4.1.2. False Positive Rates by Five Methods for False Scenario**


Proportions of Students in Level Combinations for Condition 1 of Study 1

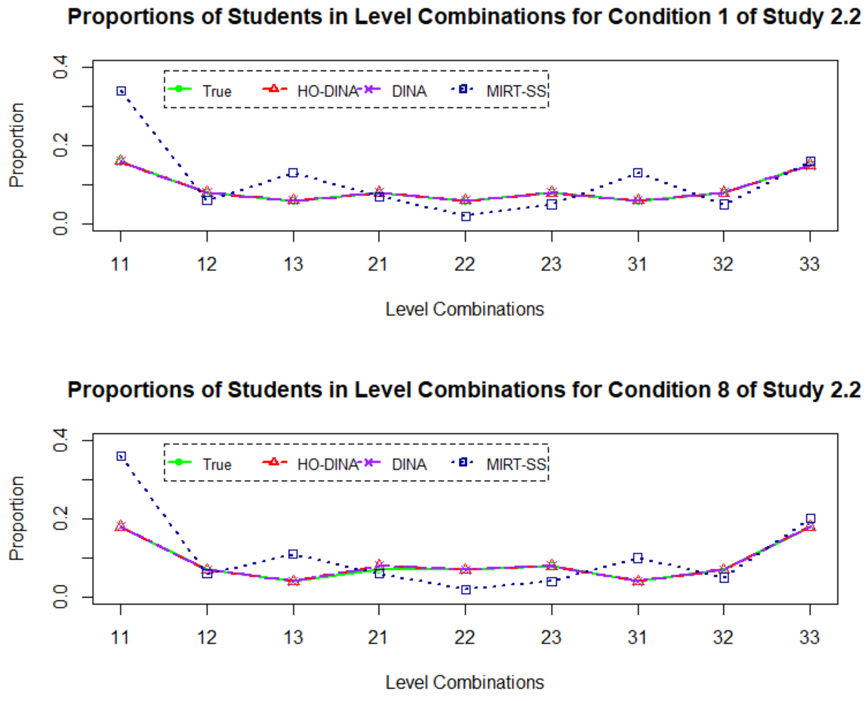
Proportions of Students in Level Combinations for Condition 8 of Study 1

Figure 4.1.3. Proportions of Students in Nine Level Links


Proportions of Students in Level Combinations for Condition 1 of Study 2.1

Proportions of Students in Level Combinations for Condition 8 of Study 2.1

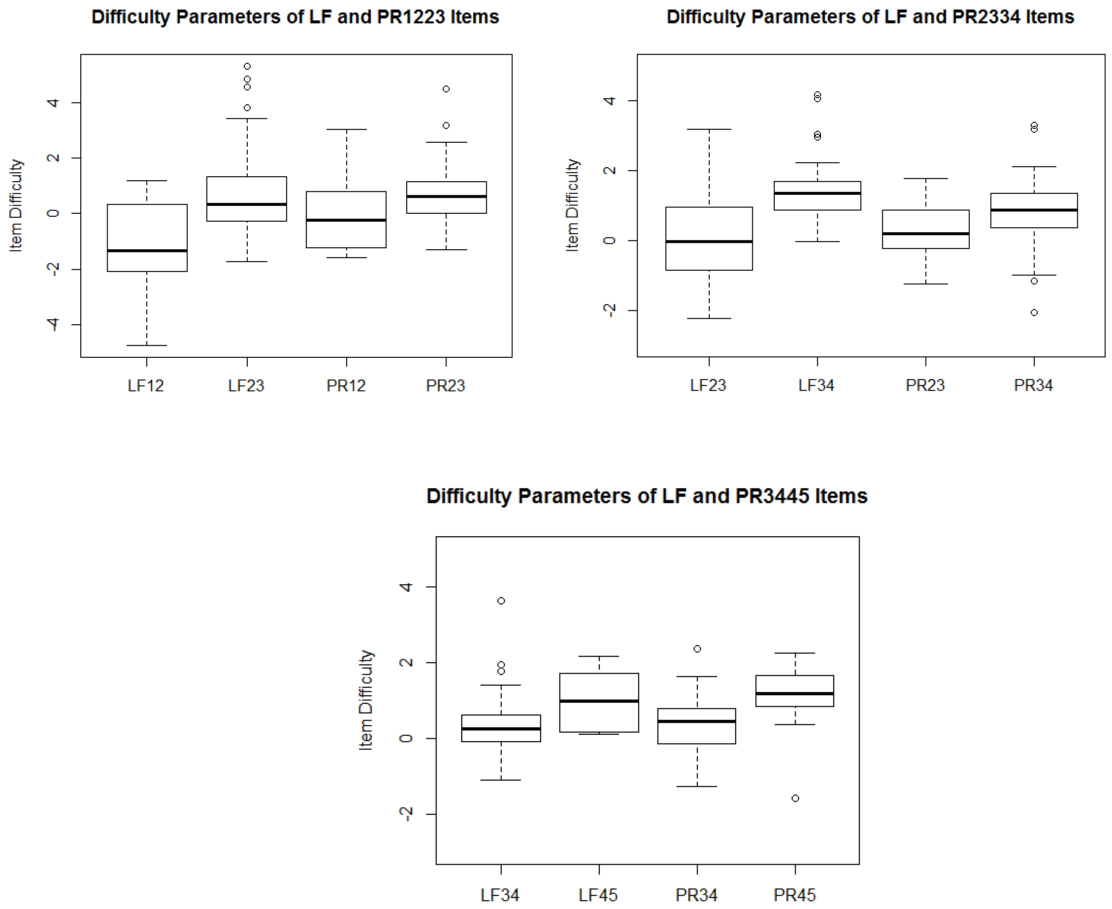
Figure 4.2.1. Proportions of Students in Level Links (Extreme Difference Cases)



Proportions of Students in Level Combinations for Condition 1 of Study 2.2

Proportions of Students in Level Combinations for Condition 8 of Study 2.2

**Figure 4.2.2. Proportions of Students in Level Links (moderate difference cases)**

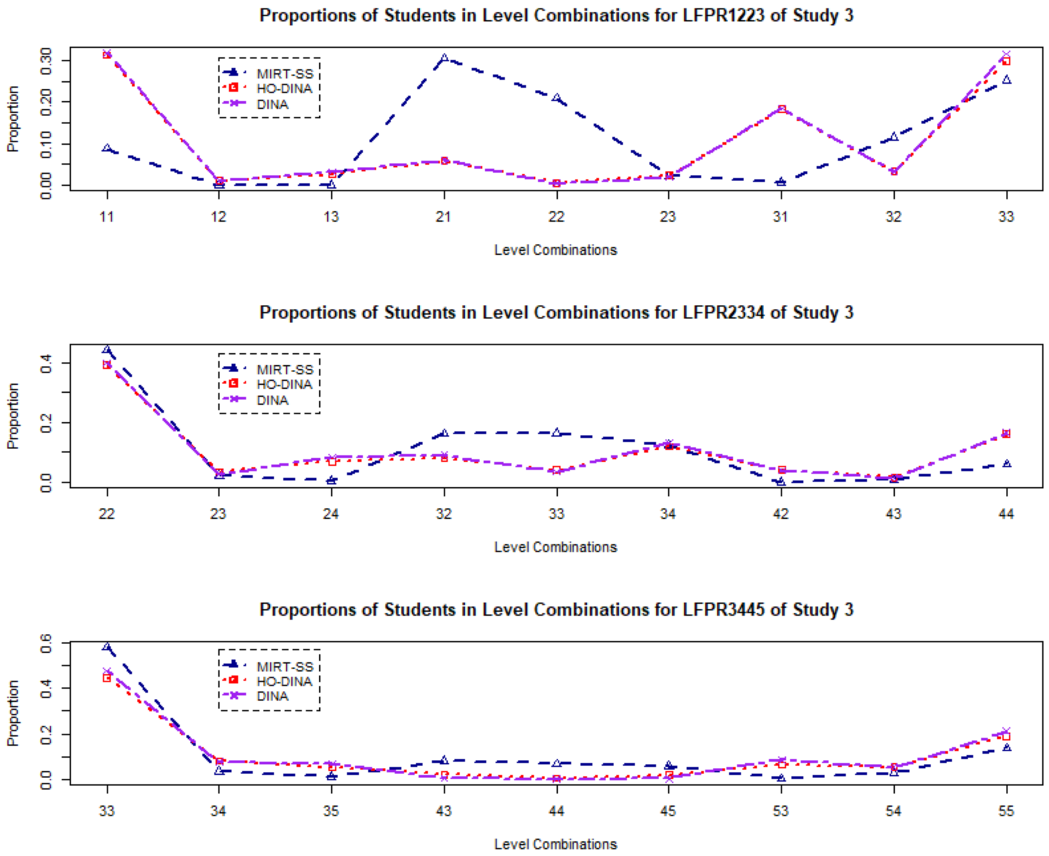


Difficulty Parameters of LF and PR1223 Items

Difficulty Parameters of LF and PR2334 Items

Difficulty Parameters of LF and PR3445 Items

**Figure 4.3.1. Ordering of Difficulty Estimates of Items Measuring Different Levels**


Proportions of Students in Level Combinations for LFPR1223 of Study 3

Proportions of Students in Level Combinations for LFPR2334 of Study 3

Proportions of Students in Level Combinations for LFPR3445 of Study 3

**Figure 4.3.2. Observed Proportions of Students in Level Combinations**

# CHAPTER V

# DISCUSSION

While learning progressions show promise and are expected by many scholars to provide granular information about student learning to support instruction and learning growth, how to empirically validate them remains a challenging problem for the field of education (Heritage, 2008; Wilson, 2012). A popular validation approach is to use statistical models to analyze response data collected from assessments developed to measure knowledge and skills specified by learning progressions. Inferences drawn from the analysis can allow us to examine theoretical claims about how the learning levels should be ordered and the plausibility of co-occurrence of the levels across progressions. Under this context, three psychometric models were investigated in this study to shed light on their effectiveness for evaluating learning progressions using simulated and empirical data. In Chapter 4, the results were reported for two simulation studies and one empirical investigation. What follows will be a discussion of the findings across the studies, with a specific connection to the research literature of evaluating learning progressions. Then, a summary of the findings about the effectiveness of the models will be provided. Finally, four limitations and a few future directions will be discussed with the intention that they will be helpful for future studies and operational works related to learning progressions.

## 5.1. Simulation Studies

When statistical models are used to evaluate learning progressions empirically, we must deal with two moving parts: (i) the trustworthiness of the learning theories, and (ii) the sensitivity of the methods using results from fitting the selected model to learning progression data. In one case, the theory can be plausible, but the model might not be sensitive enough to support the underlying theory. In another situation, learning progressions are less likely to hold, nonetheless, statistical results from fitting the model to empirical data can falsely inform the opposite due to their insensitivity of detecting implausible theories. This view was the reason why simulation Studies 1 and 2 were conducted with the purpose of understanding the moving part of model/method sensitivity in detecting the validity of learning progression theories. The results reported in Chapter 4 indicated that the simulation investigations cast some light on the effectiveness of the models and the methods using such models to examine simulated progressions. Following the organization in Chapter 4, the next two sections will discuss the results of the simulation studies by claims 1 and 2 followed by a comparison of the results across models and simulation conditions.

### 5.1.1. Claim 1: Level Order

None of the five methods considered in this study outperformed the others in terms of obtaining expected true and false positive rates at the same time. Across Studies 1 and 2, the t-test method under the MIRT framework consistently had deflated type I error rates (i.e., false positives). All the rates were below .05 and more than 80% of them did not exceed .01. This result suggests that the t-test was a strict test. In other words, it seemed to be helpful in detecting incorrectly ordered learning levels. However, it was observed across the simulation that this t-test method was not sensitive enough to detect true level order when the difference between the levels was moderate or quite large. It was, though, perfectly sensitive as any other methods when the difference between learning levels was extreme. Results for the remaining methods (i.e., ordered-test using MIRT, location using HO-DINA, and minimum tests using HO-DINA and DINA) were observed to be in the opposite direction with the t-test. Indeed, the true positive rates for these methods across true conditions were much more reasonable those of the counterpart. The lowest true positive rate was .73 for the HO-DINA location test in Study 1.

Most of the remaining rates varies from .90 to 1. Notably, the CDM-based methods perfectly reconfirmed the true information used to generate the data in Study 2 even if the difference between the attributes was moderate. This set of results signifies that the last four methods tended to be powerful enough to confirm the correct level order of learning progressions. However, these methods were less likely to perform adequately when the theory was false. In effect, their false positive rates when item difficulty and attribute locations were sampled from the same distributions were consistently much higher than a conventional error rate of .05. Across conditions, false positive rates for these methods ranged from .15 to .41 with most of the values were beyond .20. Taken together, the methods and models appeared to complement each other in detecting level order across simulation conditions. The t-test was seen to have enough power to confirm the true theory or correctly detect false progressions for conditions of 60 items. Meanwhile, the remaining methods appeared to be useful to analyze data of smaller sample sizes or less items which is very likely to be the case for classroom and/or interim assessments. In this instance, if the ordering is probable due to prior knowledge of the theory, the ordered-median, location and the minimum tests can be adopted to confirm that information. Otherwise, the t-test should be conducted to defy it.

### 5.1.2. Claim 2: Level Link

Validating the co-occurrence of levels across progressions was shown to be quite challenging even in simulation studies. As observed Study 1 when MIRT-SS was the generating model, the accuracy rates of classifying students into combinations of levels by the MIRT-SS and CDMs were far from perfect. The rate was highest at .69 in condition 8 of 1,000 students, 60 items and strong correlation for MIRT-SS and lowest at .33 in condition 5 of 1,000 students, 40 items and moderate correlation for DINA. The accuracy rates for Study 2 between MIRT-SS and the true classification by HO-DINA were higher than the first study. Nonetheless, they all remained below .84 and became as small as about .50 when the distance between attributes was set at a moderate value. These accuracy rates can be further understood by examining how the models classified students into nine reasonable combinations and seven combinations with at least one inconsistent profile in them. With the only exception when the attributes were set at extreme distance, the percentages of students in the nine combinations were seen to be notably different for MIRT-SS and the CDMs. The difference in the percentages was more likely to occur for combinations [11], [22] and [33]. In Study 1, much more students were placed by in level-2 of both progressions by MIRT-SS than by the CDMs. In Study 2, the classification accuracy and cross-model consistency into level combinations by the models depended on the magnitude of the distance between the attributes. When the distance was extreme (i.e., 2 logits), the three models appeared to work quite accurately and consistently in classifying students into levels. Expectedly, when the distance became moderate (i.e., .50 logits), the classification by the MIRT-SS and CDMs diverged greatly.

### 5.1.3. Results across Models and Conditions

In terms of how different models perform across the simulation studies, two CDMs (i.e., HO-DINA & DINA) seemed to produce comparative results across the simulation studies and their results were different from MIRT-SS' in most of the cases. Their true and false positive rates as well as proportions of students classified in each level combination were nearly identical within each condition and scenario simulated in Studies 1 and 2. Most if not all the differences by the models were within a few percentage points. In comparison to accuracy rates of around 60% to 70% of MIRT-SS in Study 1, the CDMs showed nearly-perfect recovery rates in Study 2. Indeed, the lowest classification accuracy rate for HO-DINA and DINA in the former was as high as .97. Whereas, the highest accuracy rate of MIRT-SS in Study 1 was as low as .69. This difference for the classification accuracy seemed to favor the CDMs in the sense that sampling error tended to cause less impact to the accuracy of the classification by the two models. Whereas, the use of cut scores in the MIRT framework to locate students into learning levels were more likely to be impacted by sampling error.

It is also helpful to bring into light the dependency of the result on the factors manipulated in the simulations. Looking across studies and conditions, the results appeared to be somewhat dependent on the simulated variables which included sample sizes, number of items, the strength of the correlation between progressions, and the magnitude of the distance between learning levels. And, the dependency of the results on the variables also varied by methods and models as well. Indeed, holding other factors constant, true positive rates for all five methods tended to be higher for data with more items in Study 1 when they were used to evaluate the true level order. This finding can be seen from Figure 4.1.1 in that the lines seemed to go up from conditions 1 to 4, and 5 to 8. It is noted that 40 items were generated in conditions 1, 2, 5 & 6, and 60 items in the remaining conditions. Moving to the second study, this result seemed to hold true for MIRT-based t-test for the moderate difference cases. When the difference between learning levels was extreme, the true positive rates were perfect (i.e., 1) for all methods and conditions. Coming back to the moderate difference cases, the power rates for the t-test were higher for test forms of more items. Table 4.2.2 showed us that moving from 40 items to 60 items helped increase the power rates by around 10% for conditions of 500 students. This increase in power rates for the sample size of 1,000 simulees was doubled at around 20% when more items were involved. The impact of having more items was quite salient for the MIRT order-test in the moderate difference cases. However, the influence of longer tests to the true positive rates of the order-test was much less noticeable. The rates increased at around only a few percentage points when more items were involved.

The impact of the simulated variables into the effectiveness of the models and methods in assessing the second claim of learning progressions can also be seen through the classification accuracy and cross-model consistency across conditions and studies. Overall, the accuracy and consistency rates increased slightly with the increase of sample sizes, correlation, and the number of items. For Study 1, while holding other factors constant, the stronger correlation led to a few percentage points increase in the accuracy and consistency rates. This result can be observed from Tables 4.1.6 and 4.1.7. However, the strength of the correlation did not seem to affect the accuracy and consistency rates in Study 2 (see Tables 4.2.3, 4.2.6 & 4.2.7). The accuracy rate for the CDMs appeared to depend only on the number of items for this study in which more items led to an increase of a few percentage points in the accuracy rates. The number of simulees did not seem to play a role in steering the rate up or down. Across Study 2, averaged accuracy and consistency rates for samples of 500 or 1,000 students while other variables were kept the same, were nearly identical. Across studies, five methods to detect level order worked complementarily. The accuracy rate of MIRT-SS and consistency rates between MIRT-SS and CDMs in locating students into combinations of learning levels were far from perfect which reconfirmed the challenge of evaluating the level links of learning progressions. Among the factors manipulated throughout the simulation studies, the impact of test length appeared to be the most consistent across conditions. More items led to higher true positive, accuracy and consistency rates. In brief, results of the simulation studies increased our understanding of the models and methods derived from them. It helped us draw a big picture of how the models and methods using the models functioned when one has access to the true information. This picture will guide the interpretation of the empirical results that will be discussed next.

### 5.2. Empirical Study

To investigate the effectiveness of the models and methods considered in this study in analyzing empirical data, that data were calibrated and the plausibility of the theory underlying the data was examined. Insights gained from Studies 1 and 2 about the methods and the theoretical claims became the baseline information to interpret the results of this empirical application. A summary of the key findings of the investigation and discussion of the results in refence to the theory follows.

#### 5.2.1. Claim 1: Level Order

With respect to the first claim of LR and PR, results obtained from the five methods consistently supported the theoretical ordering of learning levels for five out of six data sets. The only data that the t-test, location and minimum tests revealed negative result were LF3445. However, the order test that compared the medians of item difficulty for this data set confirmed that the medians were in an increasing order. Again, it is noted that item group of Level 4-5 of LF3445 contained only six items. In reference to the understating of how the methods worked for simulated data shown earlier, these results provided statistical evidence to support the first claim of increasing complexity of learning levels. Indeed, given the very small false positive rates of the two-sample t-test in reconfirming the level order in the true scenario, it is very likely that the lower levels of the five supported data sets (i.e., LF1223, LF2334, PR1223, PR2334 & PR3445) were less sophisticated than the higher levels. For the remaining data of LF3445, only the median test showed positive result that the median difficulty of item group LF 3-4 was smaller than the median of LF 4-5 (see Figure 4.3.1). Given the large false positive rate of this order-test and the strictness of the t-test observed throughout the simulation conditions, it is suggested that further investigation is needed to reconsider the ordering of levels 3, 4 and 5 of LF.

### 5.2.2. Claim 2: Level Link

For the second claim of co-occurrence of levels across learning progressions, using the models, students were observed in all the 10 level-links postulated in Arieli-Attali et al. (2012). Remarkably, all three models (i.e., MIRT-SS, HO-DINA, & DINA) classified at least a few students into nine out of 10 level-links. The only theorized combination that contained 0% of students was level-1 of LF and level-2 of PR. MIRT-SS did not locate any student into this link. The notable difference of the percentage of students classified into each combination of levels by MIRT-SS and CDMs shown in Tables 4.3.11 to 4.3.13 seemed to indicate that the distinctiveness of the learning levels measured by the items in this empirical study was less likely to be extremely large. Indeed, results of the extreme difference case of Study 2 signified that if the distance between attributes defined by item groups of lower and higher levels was extreme (e.g., 2 logits), the percentages of students in each combination by the models should be more similar (see Table 4.2.4). This can also be seen by looking at the location distance of the empirical data in Table 4.3.9 which varied from -.05 to .73. The moderate difference between the attributes defined by the learning levels in this empirical exploration was also likely to be the reason behind the low cross-model classification consistency among MIRT-SS and CDMs shown in Table 4.3.14. The rates ranged from .39 to .64 and were more in line with the cross-model classification consistency found in Study 2 for the moderate difference cases reported in Tables 4.2.6 and 4.2.7.

Given the evidence explained above, the models were shown to be useful in evaluating the second claim of co-occurrence of learning levels since they revealed students in all theorized combinations. On the other hand, the models also located students into seven additional links. While this finding warrants further investigation, it can be explained, in part, by looking at the nature of the theory and the additional links. For instance, combination of level-2 of LF and level-1 of PR was not postulated. However, the next combination (2,2) (i.e., level-2 for both LF and PR) was theorized and observed using each of the three models. This observation suggested that students mastered level-2 of LF might not have been at level-2 of PR automatically. It was possible that students in level-2 of the first progression can only just be competent at the knowledge and skills described in level-1 but not in the higher levels of PR. A similar argument can be made for other combinations having some students by at least one model that were not predicted by Arieli-Attali et al. (2012). In short, the models appeared to be effective in detecting the predicted order of learning levels and the possibility of co-occurrence of levels. However, how effective they are in locating students into learning levels, thus combinations of levels, remained unanswered within the scope of this statistical study. For this problem to be solved, some type of standard setting studies, classroom observations, teacher or cognitive interview must be conducted to provide external validity evidence to support the use of the model. These future directions will be elaborated next after discussing the limitations of this study.

## 5.3. Conclusion

This dissertation study was set up and implemented to examine the effectiveness of MIRT-SS, HO-DINA and DINA in evaluating two theoretical claims of learning progressions. Through two simulation studies and one empirical analysis, it can be concluded that the models and methods derived from them appeared to be effective at analyzing data to evaluate learning progressions. For the first claim of increasing order of learning levels, the MIRT t-test and the remaining four methods considered in this study were likely to work in a complementary fashion in the simulation and consistently while they were used to analyze empirical data. When the sample sizes and number of items were large enough (e.g., 1,000 students and 15 items per item group), the MIRT t-test can have a power up to .80 to confirm the true theory underlying the learning progression data. The MIRT order test, location and minimum tests of CDMs can be useful when there were less students and some prior knowledge to support the plausibility of the theory was available. In other words, these methods can be adopted to reevaluate a learning progression theory using a smaller sample and less items if this theory had been supported previously using more data and more items. This aspect of the four methods deemed useful given that a sample size of 1,000 and a test of 60 items sounds impractical for classroom or formative assessments. For this application, the very high false positive rates of the four tests would not be so concerning since our underlying theory has already been supported. Equally important, once the methods can confirm the increasing order of levels, MIRT-SS, HO-DINA or DINA could be used to identify student learning level or profile to provide information to educators to support instruction and student learning. The effectiveness of the methods to evaluate claim 1 was also seen through the empirical application. As reported in Chapter 4, results across the three models were consistent in shedding light on the increasing order of learning levels of LF and PR. Four out of five methods derived from the models which include MIRT t-test, location and minimum tests of the CDMs revealed the same test results for all data sets of the progressions (see Tables 4.3.8, 4.3.9, & 4.3.10). They all supported the increasing order of five data sets and rejected the claim for LF3445. Meanwhile, the MIRT order-test was the only test that provided evidence in favor of the theory for all the data.

In terms of the second claim of level link, Wilson (2012) described and discussed the challenge of examining this aspect of learning progression theory. Since then, at least a few studies (e.g., Pham et al., 2016; Shin et al. 2017) have tried to address this problem. The results of three studies reported earlier in this dissertation illustrated how difficult it was to evaluate the possible co-occurrence of levels across progressions when only statistical analyses were used. When MIRT-SS was the generating model, this model was able to recover only about 60% to 70% of the true classification using the true item and proficiency parameters. The accuracy this in Study 1 was improved with more data and stronger correlation between progressions but remained far from perfect. This observation indicated the challenge of using MIRT-SS and cut scores to locate students into level links. In Study 2 when the simulating model was switched to HO-DINA, the CDMs were seen to recover the true classification almost perfectly (see Tables 4.2.3 & 4.2.6). The accuracy rates for these models were 97% for conditions of moderate progression correlation and became 99% when the correlation was set at .90. These nearly perfect accuracy of the CDMs suggested that if they can fit adequately with learning progression data, the classification of students into level links by these models can be consistent enough across samples.

When assessing claim 2 in the empirical study, all three models appeared to be useful in evaluating the co-occurrence of the levels of LF and PR. Results obtained from fitting the models to the data provided evidence to support all 10 combinations predicted by the theory. They also suggested an addition of seven more possible links that could be considered to revise and reevaluate the theory. It is also noted that the cross-model classification consistency between MIRT-SS and the CDMs varied from as low as 39% for LFPR1223 to as high as 64% for LFPR3445. Since both IRT and CDMs can fit adequately with a given data set (Haertel, 1990), the difference in how these models classified students into learning levels suggested that further studies are needed to shed more light on the validity of the classification. These future directions will be discussed after the section on some limitations of this dissertation study which will come next.

### 5.3.1. Limitations

As any other scientific research, three studies reported in this dissertation have limitations. First, for practical purpose, only three specific models were considered in this study. Traditionally, CTT and IRT models has been suggested and used to evaluate learning hierarchies and progressions (e.g., Heritage, 2008; Steedle & Shavelson, 2009; White, 1974). CDMs came along later and have been adopted to explore learning progression data (e.g., Chen et al., 2017; Kizil, 2015; Pham et al., 2017). It is also noted that the Rasch model and its multi-dimensional extensions have been the main tool that was used extensively by the published works that relied on the IRT framework to evaluate learning progressions. For the CDMs, various models were adopted in the context of learning progressions. Kizil (2015) used the attribute hierarchy model by Gierl et al. (2006) and generalized diagnostic models by von Davier (2005). These models are more complicated and general than the CDMs considered in this dissertation. Similarly, Chen et al. (2017) used Rule Space Model (Tatsuoka, 1983) which is an CDM that assumes a hierarchical relationship for the attributes defined by their learning progressions. In this study, 2PL MIRT model with simple structure, the DINA and its higher-order version HO-DINA were examined instead of the Rasch model and the more generalized or hierarchy version of CDMs. Within the MIRT-SS, the cut scores to place students into learning levels were the median of item difficulty of each item group. These scores can be set differently by different methods such as using the domain characteristic curve and an appropriate response probability. Within the

CDMs, the default calibration settings by GDINA were adopted. These settings can be adjusted. Future investigations can consider a wider range of models and calibration options or even compare the Rasch framework with its counterparts of 2PL IRT.

Second, only learning progressions of three learning levels were simulated and considered in this study. Even if three level progressions are popular in the literature (Shin et al., 2017), existing progressions can have as many as 15 levels (e.g., Briggs et al., 2015). How the models perform in evaluating progressions of more than three levels remains an open question. This limitation can be addressed in follow-up studies by considering progressions with more than three levels.

Third, this set of studies only focused on the statistical aspect of evaluating learning progressions. While using statistical models deemed useful to examine learning claims empirically, validity evidence collected from other informants such as classroom teachers, content experts, or students through cognitive interviews is needed to draw more holistic view of the theory under evaluation. The qualitative information if it becomes available, can be used to interpret or critique the statistical results. For example, expert and teacher opinions can provide insightful explanation for the plausibility of level links (2,1) and (3,2) for LF and PR which were not predicted but many students were observed by MIRT-SS to be in these combinations.

Fourth, the empirical investigation in Study 3 was purely sectional in the sense that student learning was only captured at one point in time. It would also be useful if data of student learning can be tracked longitudinally to see if students transition from one level to the next as the theory predicted. Since learning progressions are really about common pathways that a typical student would go through when s/he learns a content area, knowing more about how students' progress from one level to the next through a course of study is needed to cast more light on the validity of the theory underlying LF and PR. To address the limitations, a few future directions are suggested next.

### 5.3.2. Future Directions

Three studies reported in this dissertation contributed useful information about the effectiveness of the MIRT-SS, HO-DINA and DINA in evaluating learning progressions. Nonetheless, by no mean can they answer every research question and offer final solutions to the challenge of assessing level links put forth in Wilson (2012). Under this view, a few lines of follow-up research are suggested from what was learned through this study to expand our understanding and tool kits to evaluate learning progressions. To address the first limitation described earlier, two new simulation directions can be taken. In the first place, a new set of simulations should be conducted to compare the performance of MIRT-SS, HO-DINA and DINA when adjustments are made with respect to the method used to identify cut scores or calibration features set in GDINA to estimate the CDMs. Another direction would be expanding the scope of Studies 1 and 2 to consider more statistical models. Within an IRT framework, Rasch-based approaches such as the change-point model introduced in Shin et al. (2017) can be compared with the 2PL counterparts. Similarly, a simulation study that takes into consideration some more complicated models from the CDM family can be useful to help the field understand more about how these models perform in the best-case scenario where one knows the true information of the simulated progressions. Among the CDMs, models that assume a hierarchical order of learning attributes seem to be relevant to the work of evaluating learning progressions of more than three levels. Different models should also be fit to empirical data and model-data fit information should be used to select the most appropriate model or to examine the usefulness of the model. A prior example of this line of research can be found in Kizil (2015).

Second, to consider progressions of more than three learning levels, they can be generated and investigated using the models considered in this study or some other models mentioned earlier. In this case, more cut scores will be needed and more settings for CDMs can be selected to analyze the data. The introduction of more than two cuts and a wider selection of CDMs can make the exploration more challenging. However, these follow-up investigations are expected to bring us closer to the real complexity of evaluating learning progressions of more than three levels.

Third, while simulations enable us to understand the statistical models and methods, qualitative perspectives can offer valuable insight into how learning progressions play out in teacher professional development, classroom instruction and assessment. This view suggests a line of research that follows the principle of research practice partnership (RPP) (Fishman, Penuel, Allen, Cheng, & Sabelli, 2013) to bring researchers and practitioners together to collaborate in educational research to support student learning. Following this RPP method, learning scientists, curriculum experts and teachers can work together to grasp existing learning progressions or define new theories. Then, teachers will rely on the progressions to design lessons, activities and build classroom assessment. In the next step, teachers implement the curriculum and work with researchers to collect and analyze classroom data and student artifacts. These data will be used to evaluate the progressions and revise the theory. This process can be looped in cycles as a continuous improvement tool to refine the learning theory.

Fourth, collecting longitudinal data to evaluate LF and PR would advance the study of these progressions into another level. Items can be subset from the current item pools for LF and PR. For some levels (e.g., Level 1-2 and Level 4-5 for LF and PR), more items can be drafted and revised to add in the existing pools. In the next step, test forms will be built and administered to the same group of students at multiple time points. If resources become available, external variables of student learning such as their math scores or self-confidence ratings can be gathered to provide evidence to validate the test scores of the participants. Longitudinal IRT or CDMs can be adopted to analyze the longitudinal data. Other sources of information such as classroom videos, students' worksheets and artifacts can also be collected and analyzed to shed more light on the learning trajectories through which each student advances their knowledge and sharpen their skills of functions, linear functions and proportional reasoning from novice to expert understanding and expertise.

Fifth, given that computers are much more accessible to students nowadays than a few decades ago, the line of research on using learning hierarchies to improve computer-based testing initiated by Ferguson (1969) and advanced by Spineti and Hambleton (1977) should be revived. Recently, adaptive assessment systems have been developed to take advantage of learning maps and trajectories to support personalized learning (e.g., Confrey et al., 2017) and students of special needs (e.g., Dynamic Learning Maps® Consortium, 2018). However, more studies along this line should be conducted to build more knowledge around this topic and inform developmental projects and useful applications of learning progressions and assessments based on learning theories.

Finally, to improve the utility of assessments based on learning progressions, it is critical that the communication of information obtained from the instrument to different stakeholders need to be effective and useful. In other words, studies of how to report assessment results regarding the current learning profile of learners and provide feedback to them and educators should be carried out. For instance, these studies can inform the kind of learning report layouts and presentations that might be the most accessible and useful for students, teachers and parents.

Taking the LF as an example, a few graphs of linear functions of different slopes plotted in the same coordinate plane can be used in a learning report as a suggestion for the next learning step for a student in learning level-3 of this progression. It is reminded that level-3 students in LF can understand and can work well with one linear function. In short, six follow-up directions were suggested in the hope that if they are carried out to a certain extent a more comprehensive view of the learning progression landscape will take shape. What comes next is some take-away messages that were drawn from this study to share with researchers and practitioners who are interested in evaluating learning progressions.

### 5.3.3. Practical Implications

In this last section, three practical implications will be discussed as a way to conclude the dissertation. The first implication is about model selection. Then, the second set of suggestions deal with how to do data collection to evaluate learning progressions effectively. Last, collecting different sources of evidence to validate learning progressions will be elaborated as a take-away message for the audience.

As reported in the literature review, IRT and CDM are the two main modeling frameworks that have been used to evaluate learning progressions. Given the difference of the results for MIRT-SS, HO-DINA and DINA found in this study in many conditions as well as the empirical analyses, it can be generalized that IRT models and CDMs do not necessarily result in consistent classification of students into learning levels. Thus, the choice of model from one framework over the other is needed to provide more useful and valid information of student learning and feedback for instructional purpose. To avoid issues related to retrofitting, models should be selected prior to assessment development and data collection. As discussed in the literature review, IRT assumes the continuity of the construct measured by the assessment.

Whereas, the construct under CDM is assumed to be discrete. This distinctiveness of the frameworks implied that model selection should hinge on the nature of the construct defined by learning progressions and their learning levels. On the one hand, if the levels involve simple knowledge and skills such as adding two single-digit numbers, they can be dichotomized into mastery or non-mastery and CDMs can be preferred over IRT. On the other hand, if the levels appear to be spread out and cover a range of related yet different concepts and skills such as understanding and being able to work with non-linear functions, an IRT model seems to be a more appropriate choice.  Once the model is chosen, they can be used to inform item development, data collection and data analysis.

The second line of implications relates to data collection to evaluate learning progressions under consideration. In the best-case scenario, statistical models should be selected before the construction of assessment and data collection. If it is the case, simulation studies should be conducted to guide the data collection design and analysis plan. The missingness by design of the empirical data used in Study 3 made it more challenging to examine IRT assumptions and model-data fit for all the models considered in this study. To mitigate the possible challenges caused by missingness or lack of power due to sample sizes, simulations can be carried out to compare a few data collection options to inform the most adequate design. Studies 1 and 2 are  two examples of how to use simulations to inform assessment development and data collection. If MIRT-SS is adopted to evaluate learning progressions and MIRT-based t-test method is used to evaluate level order, it is suggested that at least 15 items per item group that tap into knowledge and skills of each adjacent levels are needed. If fewer items are used, the power to confirm the correct order of learning levels if it is true might be as low as .40. Even if the suggestion of 15 items per item group is fulfilled, the possible correlation between progressions is likely to play a role in steering the power up or down. The power line of MIRT-based t-test in Figure 4.1.1 implied that if the expected correlation is around .60 to .70, at least 1,000 students will be desired to keep the power in the range of .70 to .80. From what was observed from Studies 1 and 2, sample sizes and number of items did not seem to impact the results of HO-DINA and DINA. With this observation, data requirements for these models should follow conventional guidelines and sample size recommendations for CDMs (e.g., Choi, Templin, Cohen, & Atwood, 2010; Kunina-Habenicht, Rupp, & Wilhelm, 2012) or simulation studies can be conducted to inform data collection design.

Finally, according to George E. P. Box, “Essentially, all models are wrong, but some are useful.” (p.424, Box & Draper, 1987). If this view is well taken, then the responsibility of researchers using modeling as a tool to evaluate substantive learning theories begins by examining which models are less wrong and more useful. This may, however, be easier said than done, and the statistical explorations conducted in this dissertation illustrated how challenging it is to investigate the effectiveness and usefulness of only three models. Model-data fit analyses probably revealed some evidence to know which models were less wrong. Knowing which ones are more useful seems to be much more laborious and arduous. However, the finding that results from using MIRT-SS and CDMs to analyze learning progression data were quite different across numerous simulation conditions and in the empirical study implied that evidence from sources other than the internal structure of the response data is much needed to interpret the statistical results obtained from fitting the models and shed light on the usefulness of the models. As elaborated in the previous section on future directions, researchers and practitioners in the field of learning progressions should collect information from content experts, teachers, students and classroom activities to provide a more comprehensive view of the learning progressions under evaluation. It is believed that looking into student learning from multiple angles and contexts will enable us to figure out an improved way to describe, evaluate and refine learning theories for the purpose of advancing human learning and thus human conditions.

# APPENDIX: PERCENTAGES OF STUDENTS IN EACH PROFILE

In this appendix, the expected percentages of students sampled from a standard normal distribution who mastered attribute 1 (i.e., of profile [10]), or mastered both attribute 1 and attribute 2 (i.e., in profile [11]), or were in the inconsistent cognitive profile [01] in the true scenario were computed. From equation (7) in Chapter 2, one has the probability of a student of continuous proficiency $\theta$ to master attribute 1 is:

$$P(a = 1|\theta) = \frac{1}{1+\exp(-1.7(1+\theta))}.$$

And, the probability for her/him to master attribute 2 is:

$$P(a = 2|\theta) = \frac{1}{1+\exp(-1.7(-1+\theta))}.$$

Conditional on the continuous proficiency, the probability for her/him to master both attributes, thus in learning level-3, is:

$$P(a_1 = 1 \,\&\, a_2 = 1\ |\theta) = \frac{1}{(1+\exp(-1.7(-1+\theta)))*(1+\exp\left(-1.7(1+\theta)\right))}.$$

Similarly, the probability for her/him to master only attribute 2 but not attribute 1 is:

$$P(a_1 = 0 \,\&\, a_2 = 1\ |\theta) = \left(1 - \frac{1}{1+\exp(-1.7(1+\theta))}\right) * \frac{1}{(1+\exp\left(-1.7(1+\theta)\right))}.$$

In general, if the probability of mastering attribute 1, or both of them, or only one of them is given, let's denote it be *P (. | $\theta$).* Then, for a population of examinees of a certain proficiency distribution with a density function of *f($\theta$),* the overall percentage of students in this population mastering attribute 1, or both of them, or only one of them is computed as the integral over the whole range of $\theta$ of the product of *P (. | $\theta$)* and *f($\theta$).* In functional form, it can be written as:

$$p_{+}(.) = \int_{-\infty}^{+\infty} P\ (.\,|\,\theta) * f(\theta) * d\theta.$$

To approximate the percent correct for mastering attribute 1, or both of them, or only attribute 1, the following R-codes were used.

```
D=1.7 # to set up the normal-ogive scale for the attribute
itcp1 <- .25 # -1
itcp2 <- -.25 # -1
integrand1 <- function(x){1/(1+exp(-D*(itcp1+x)))}                    # for mastering attribute 1
integrand2 <- function(x){1/(1+exp(-D*(itcp2+x)))}                   # for mastering attribute 2
integrand12 <- function(x){1/((1+exp(-D*(itcp1+x)))*(1+exp(-D*(itcp2+x))))} # for mastering
both attribute 1 & 2
integrand01 <- function(x){(1-1/(1+exp(-D*(itcp1+x))))*(1/(1+exp(-D*(itcp2+x))))} # for
mastering attribute 2 but not  1


product1 <- function(x){dnorm(x,0,1)*integrand1(x)}   # for mastering attribute 1
product2 <- function(x){dnorm(x,0,1)*integrand2(x)}   # for mastering attribute 2
product12 <- function(x){dnorm(x,0,1)*integrand12(x)} # for mastering both attribute 1 & 2
product01 <- function(x){dnorm(x,0,1)*integrand01(x)} # for mastering attribute 2 but not 1

integrate(product1,lower=-6,upper=6)   # for mastering attribute 1
integrate(product2,lower=-6,upper=6)   # for mastering attribute 2
integrate(product12,lower=-6,upper=6)  # for mastering both attribute 1 & 2
integrate(product01,lower=-6,upper=6)  # for mastering attribute 2 but not 1
```

Below is the results I obtained when I ran the code.

```
# For extreme difference cases: itcp1 <- 1, itcp2 <- -1

      >integrate(product1,lower=-6,upper=6)   # for mastering attribute 1
      .7592567 with absolute error < 6.4e-05

      >integrate(product2,lower=-6,upper=6)   # for mastering attribute 2
      .2407433 with absolute error < 8.6e-05

      >integrate(product01,lower=-6,upper=6)  # for mastering attribute 2 but
      not 1
      .01790194 with absolute error < 4.7e-06

# For moderate difference cases: itcp1 <- .25, itcp2 <- -.25

      > integrate(product1, lower=-6,upper=6)   # for mastering attribute 1
      .5701569 with absolute error < 6.1e-06

      > integrate(product2, lower=-6,upper=6)   # for mastering attribute 2
      .4298431 with absolute error < 4.5e-06

      > integrate(product01, lower=-6,upper=6)  # for mastering only attr. 2
```

# REFERENCES

American Educational Research Association, American Psychological Association, & National Council for Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Arieli-Attali, M., Wylie, E.C., Bauer, M. I. (2012, April). *The use of three learning progressions in supporting formative assessment in middle school mathematics.* Paper presented at the 74th annual meeting of the American Educational Research Association, Vancouver, Canada.

Bailey, A. L., & Heritage, M. (Eds.). (2008). *Formative assessment for literacy, grades K-6: Building reading and academic language skills across the curriculum*. Thousand Oaks: CA, Corwin Press.

Black, P., Wilson, M., & Yao, S. Y. (2011). Road maps for learning: A guide to the navigation of learning progressions. *Measurement: Interdisciplinary Research & Perspective*, *9*(2-3), 71-123.

Briggs, D. C., Alonzo, A. C., Schwab, C., & Wilson, M. (2006). Diagnostic assessment with ordered multiple-choice items. *Educational Assessment*, *11*(1), 33-63.

Briggs, D. C., Diaz-Bilello, E., Peck, F., Alzen, J., Chattergoon, R., & Johnson, R. (2015). *Using a learning progression framework to assess and evaluate student growth*. National Center for the Improvement of Educational Assessment Inc. (NCIEA): Dover, NH.

Briggs, D. C., & Peck, F. A. (2015). Using learning progressions to design vertical scales that support coherent inferences about student growth. *Measurement: Interdisciplinary Research and Perspectives*, *13*(2), 75-99.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443-459.

Box, G. E., & Draper, N. R. (1987). *Empirical model-building and response surfaces*. New York: John Wiley & Sons.

Cai, L. (2010). A two-tier full-information item factor analysis model with applications.  *Psychometrika*, *75*, 581-612.

Cai, L. (2015). flexMIRT® version 3: Flexible multilevel multidimensional item analysis and test scoring [Computer software]. Chapel Hill, NC: Vector Psychometric Group.

Camara, W., O'Connor, R., Mattern, K., & Hanson, M. A. (2015). *Beyond academics: A holistic framework for enhancing education and workplace success* (Research Report No. 04-15). Iowa City, Iowa: ACT Inc.

Chen, F., Zhang, S., Guo, Y., & Xin, T. (2017). Applying the Rule Space Model to develop a learning progression for thermochemistry. *Research in Science Education, 47,* 1357-1378.

Chen, J. (2012). *Applying item response theory methods to design a learning progression-based science assessment* (Doctoral dissertation). Retrieved from ProQuest LLC. (Accession No. 1-267-18583-X)

Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in Cognitive Diagnosis Modeling. *Journal of Educational Measurement, 50,* 123-140.

Chen, W. H., & Thissen, D. (1997). Local dependence indices for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22,* 265-289.

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

Confrey, J., Gianopulos, G., McGowan, W., Shah, M., & Belcher, M. (2017). Scaffolding learner-centered curricular coherence using learning maps and diagnostic assessments designed around mathematics learning trajectories. *ZDM Mathematics Education*, *49*(5), 1-18.

Confrey, J., Jones, R. S., & Gianopulos, G. (2015). Challenges in modeling and measuring learning trajectories. *Measurement: Interdisciplinary Research and Perspectives, 13*, 100-105.

Confrey, J., Maloney, A. P., & Corley, A. K. (2014). Learning trajectories: A framework for connecting standards with curriculum. *ZDM Mathematics Education*, 46, 719-733.

Corcoran, T., Mosher, F. A., & Rogat, A. (2009). *Learning progressions in science: An evidence-based approach to reform*. New York, NY: Center on Continuous Instructional Improvement, Teachers College - Columbia University.

Daro, P., Mosher, F., & Corcoran, T. (2011). *Learning trajectories in mathematics: A foundation for standards, curriculum, assessment, and instruction*. (Report No. RR-68). Philadelphia, PA: Consortium for Policy Research in Education.

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

Dorans, N. J., & Lawrence, I. M. (1987). *The internal construct validity of the SAT* (Report No. RR-35-87). Princeton, NJ: Educational Testing Service.

Dynamic Learning Maps® Consortium. (2018, January). *2016–2017 Technical Manual Update – Science*. Lawrence, KS: University of Kansas, Center for Accessible Teaching, Learning, and Assessment Systems (ATLAS).

Erlwanger, S. (1973). Benny's conceptions of rules and answers in IPI mathematics. *Journal of Children's Mathematical Behavior, 1,* 7-26.

Faure, E., Herrera, F., Kaddoura, A., Lopes, H., Petrovsky, A. V., Rahnema, M., & Ward F. C. (1972). *Learning to be: The world of education today and tomorrow.* Paris: UNESCO.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*, 1149-1160.

Ferguson, R. L. (1969). *The development, implementation, and evaluation of a computer-assisted branched test for a program of individually prescribed instruction* (Unpublished doctoral dissertation). University of Pittsburgh, Pittsburgh, PA.

Frey, M. C., & Detterman, D. K. (2004). Scholastic assessment or g? The relationship between the scholastic assessment test and general cognitive ability. *Psychological Science, 15,* 373-378.

Fu, J., Chung, S., & Wise, M. (2013). *Dimensionality analysis of CBAL™ writing tests* (Report No. RM-13-01). Princeton, NJ: Educational Testing Service.

Furtak, E., Morrison, D., & Kroog, H. (2014). Investigating the link between learning progressions and classroom assessment. *Science Education*, *98*, 640-673.

Gagne, R. M. (1962). The acquisition of knowledge. *Psychological Review, 69*(4), 355.

Gierl, M. J., Leighton, J. P., & Hunka, S. (2007). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and practices*. Cambridge University Press.

Greaney, K., & Tunmer, W. E. (2010). The literacy learning progressions and the reading and writing standards: Some critical issues. *Kairaranga*, *11*(2), 23-27.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review*, *9*(2), 139-15.

Haertel, E. H. (1990). Continuous and discrete latent structure models for item response data. *Psychometrika, 55(3)*, 477-494.

Haertel, E., Beauregard, R., Confrey, J., Gomez, L., Gong, B., Ho, A. D., ... & Shepard, L. (2012). NAEP: Looking ahead. Leading assessment into the future. *National Center for Education Statistics. Initiative on the Future of NAEP. Washington, DC*.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational measurement: issues and practice*, *12*(3), 38-47.

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433-470). Westport, CT: Praeger.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Heritage, M. (2008). *Learning progressions: Supporting instruction and formative assessment*. Washington, DC: Chief Council of State School Officers (CCSSO).

Houwer, J. D., Barnes-Holmes, D., & Moors, A. (2013). What is learning? On the nature and merits of a functional definition of learning. *Psychonomic Bulletin & Review, 20(4)*, 631-642.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*, 527-535.

Kane, M. T., & Bejar, I. I. (2014). Cognitive frameworks for assessment, teaching, and learning: A validity perspective. *Psicología Educativa, 20(2)*, 117-123.

Kingston, N. M., Broaddus, A., & Lao, H. (2015). Some thoughts on "Using learning progressions to design vertical scales that support coherent inferences about student growth". *Measurement: Interdisciplinary Research and Perspectives, 13*(3-4), 195-199.

Kizil, R. C. (2015). *The marginal edge of learning progressions and modeling: Investigating diagnostic inferences from learning progressions assessment* (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global (Accession No. 3743727).

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *Journal of Educational Measurement*, *49*(1), 59-81.

Lee, Y. S., de la Torre, J., & Park, Y. S. (2012). Relationships between cognitive diagnosis, CTT, and IRT indices: an empirical investigation. *Asia Pacific Education Review*, *13*(2), 333-345.

Lobato, J., & Walters, C. D. (2017). A taxonomy of approaches to learning trajectories and progressions. In J. Cai (Ed.), *Compendium for research in mathematics education (*pp. 74-101). Reston, VA: National Council of Teachers of Mathematics.

Lord, F. M., & Novick, M. R. (1980). *Statistical theories of mental test scores*. Reading, MA: Adison-Wesley.

Ma, W. & de la Torre, J. (2017). GDINA: The generalized DINA model framework. R package version 1.4.2. Retrieved from https://CRAN.R-project.org/package=GDINA.

Martone, A., & Sireci, S. G. (2009). Evaluating alignment between curriculum, assessment, and instruction. *Review of Educational Research, 79(4)*, 1332-1361.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149-174.

Masters, G., & Forster, M. (1996). *Developmental assessment: Assessment resource kit.* Camberwell, Vic: The Australian Council for Educational Research.

Mislevy, R.J. (1996). Test theory reconceived. *Journal of Educational Measurement, 33,* 379-416.

Neumann, K., Viering, T., Boone, W. J., & Fischer, H. E. (2013). Towards a learning progression of energy. *Journal of Research in Science Teaching*, *50,* 162–188.

Paik, S., Song, G., Kim, S., & Ha, M. (2017). Developing a four-level learning progression and sssessment for the concept of buoyancy. *Eurasia Journal of Mathematics, Science and Technology Education*, *13*(8), 4965-4986.

Pellegrino, J., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. Committee on the Foundations of Assessment, National Research Council. Washington, D.C.: National Academy Press.

Pham, D. N., Bauer, M. I., Wylie, C., & Wells, C.S. (2017, October). *Using cognitive diagnosis models to evaluate a learning progression theory*. Paper presented at the annual meeting of Northeastern Educational Research Association. Trumbull, CT.

Pham, D. N., Monroe, S., & Wells, C. S. (2016). *Evaluation of learning progression for linear functions, proportional reasoning, and equality and variable* (Report No. 936). Amherst, MA: Center for Educational Assessment.

Piaget, J., & Inhelder, B. (1975). *The origin of the idea of chance in children*. New York, NY: W. W Norton.

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/.

Rasch, G. (1960). *Probabilistic models for some attainment and intelligence tests*. Copenhagen: Danish Institute for Educational Research

Reckase, M. (2009). *Multidimensional item response theory (Vol. 150).* New York: Springer.

Resnick, L. B. (1973). Hierarchies in children's learning: A symposium. *Instructional Science, 2*(3), 311-361.

Robin, F., Bejar, I., Liang, L., & Rijmen, F. (2016). Dimensionality analyses of the GRE® revised General Test verbal and quantitative measures (Report No. RR-16-20). Princeton, NJ: Educational Testing Service.

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

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

Shin, H. J., Wilson, M., & Choi, I. H. (2017). Structured constructs models based on change point analysis. *Journal of Educational Measurement, 54*(3), 306-332.

Simon, M. A. (1995). Reconstructing mathematics pedagogy from a constructivist perspective. *Journal for Research in Mathematics Education, 26,* 114–145.

Sinharay, S., & Haberman, S. J. (2014). How often is the misfit of item response theory models practically significant? *Educational Measurement: Issues and Practice, 33*(1), 23-35.

Spineti, J. P., & Hambleton, R. K. (1977). A computer simulation study of tailored testing strategies for objective-based instructional programs. *Educational and Psychological Measurement*, *37*(1), 139-158.

Steedle, J. T., & Shavelson, R. J. (2009). Supporting valid interpretations of learning progression level diagnoses. *Journal of Research in Science Teaching, 46*(6), 699-715.

Swaminathan, H., Hambleton, R. K., & Rogers, H. J. (2007). Assessing fit in item response Models. In  C. R. Rao, & S. Sinharay (Eds.), *Handbook of Statistics: Psychometrics* (pp. 683-718). London: Elsevier Publishing Co.  Double check the format here.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*(3), 393-408.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, *20*(4), 345-354.

Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconceptions by the pattern classification approach. *Journal of Educational Statistics, 10*(1), 55-73.

Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In R. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 327-359). Hillsdale, NJ: Erlbaum.

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

Thissen, D. (2015). Growth through levels. *Measurement: Interdisciplinary Research and Perspectives, 13*(2), 128-131.

von Davier, M. (2005). *A general diagnostic model applied to language testing* data (Report No. RR-05-16). Princeton, NJ: Educational Testing Service.

White, R. T. (1973). Research into learning hierarchies. *Review of Educational Research*, *43*(3), 361-375.

White, R. T. (1974). The validation of a learning hierarchy. *American Educational Research Journal*, *11*(2), 121-136.

Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching*, *46*(6), 716-73.

Wilson, M. R. (2012). Responding to a challenge that learning progressions pose to measurement practice. In A. C. Alonzo & A. W. Gotwals (Eds.), *Learning progressions in science: Current challenges and future directions*, 317–343. Rotterdam, The Netherlands: Sense Publishers.

Wilson, M. R., & Bertenthal, M. W. (Eds.). (2005*). Systems for state science assessment*. Washington DC, DC: The National Academies Press.

Winkles, J. (1986). Achievement, understanding, and transfer in a learning hierarchy. *American Educational Research Journal*, *23*(2), 275-288.

Wright, B. D., & Douglas, G. A. (1975). *Best test design and self-tailored testing*. Statistical Laboratory, Department of Education, University of Chicago.

Wylie, E.C, Arieli-Attali, M., Bauer, M. I. (2014, April). *Channeling teacher noticing with learning progression-based formative assessment*. Paper presented at the 76th annual meeting of the American Educational Research Association. Philadelphia, PA.

Zwick, R. (1987). Assessing the dimensionality of NAEP reading data. *Journal of Educational Measurement*, *24*(4), 293-308.