Detection of Attribute Hierarchies and Classification Accuracy: the Value of the Hierarchical Diagnostic Classification Model in Formative Assessment Practices

By

Copyright 2017

Linette Mar'ea McJunkin

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

Chairperson Dr. John Poggio

Dr. Jonathan Templin

Dr. Susan Embretson

Dr. William Skorupski

Dr. Kelli Thomas

Dr. Angela Broaddus

Date Defended: 20 February 2017

The Dissertation Committee for Linette M. McJunkin

certifies that this is the approved version of the following dissertation:

Detection of Attribute Hierarchies and Classification Accuracy: the Value of the

Hierarchical Diagnostic Classification Model in Formative Assessment Practices

Chairperson Dr. John Poggio

Date Approved: 20 March 2017

Abstract

The assumption that learning occurs sequentially, or in steps, is a common consideration in K12 education. As the landscape of student education continues to advance, driven by efforts to incorporate tools that offer educators and students feedback capable of identifying areas a student is excelling or struggling in, cognitive diagnostic models are emerging as potentially effective and efficient tools. Despite the value of diagnostic models, there are concerns regarding the application of these models when as learning hierarchy is present or theorized; applying nonhierarchical cognitive diagnostic models when an attribute hierarchy is present or applying hierarchical cognitive diagnostic models when an attribute hierarchy is not present influences the classification accuracy of the models. This study was designed to evaluate the efficiency of the HDCM in statistically testing for an attribute hierarchy, therein providing researchers with evidence and support for subsequent model application. By evaluating the results from a formative assessment designed based on the structure of a theorized attribute hierarchy, this study highlighted the model fit variations and student classification differences. The results of this study indicate that the HDCM does in fact provide a means for researchers to investigate the presence of a theorized hierarchy. Additionally, this study highlights the potential classification differences noted between models.

Acknowledgments

I am grateful to each of my committee members for their knowledge, expertise, guidance, and continued support, as each has played a significant role in the completion of this work. Dr. Poggio, the constant thinker, has challenged me to never accept status quo, taught me the value of networking, and has encouraged me to always ask questions. His passions for learning, formative assessment, and innovation have left an indelible mark that will continue to influence my career, research and life. Dr. Templin's dedication and his understanding of diagnostic classification models has broadened my view of diagnostic assessment, formative assessment and the impact proper analysis can have on student learning. Dr. Embretson's ability to clearly explain challenging concepts has pushed me to consider alternative solutions in every situation. Dr. Skorupski's enthusiasm for understanding transcends teaching measurement; his drive to absorb, evaluate, and comprehend has influenced how I approach unfamiliar concepts, pressing me to not only consider comprehension but means of application as well. Dr. Thomas has helped me think purposefully about student reasoning and learning, allowing me to evaluate the capacity of formative assessment. Dr. Broaddus has been a true support; her understanding of student learning and the needs of teachers has been irreplaceable to me during this study.

To my family, I am eternally grateful for the love and support. I cannot begin to adequately express my gratitude to my wife Kay, as she has seen me through the most challenging days; with confidence and encouragement she continues to remind me to acknowledge my abilities. Her love and never-ending support have, and always will be my true north. My mother, my warrior, has served as an inspiration to me throughout my life; showing me the value of educators in student learning. Her passion for children and commitment to those she loves will always guide me, personally and professionally. To my children Danielle and

Grant, the constants in my heart, who have taught me about patience, commitment and perseverance; I hope to have shown them that they are the joys of my life and that there is nothing more important than family. To Jaryl, for giving me a view of the world through kindness and joy, one uncluttered by greed and trepidation; I am so grateful for the pure love and blind support.

I am indebted to Dr. Gwen Carnes for igniting a passion in me that formulated my entire graduate path and professional career, for always encouraged me to continue learning. To Sarah McConnell, my trusted friend, for her words of encouragement, unwavering support (even in procrastination), and shared comradery. And finally, to the many other friends and colleagues who have supported me with encouragement, patience and kindness, thank you.

Table of Contents

CHAPTER 1

Introduction

There is determination in education to provide quality instruction with assessment practices designed to elicit expansive information about student learning and understanding. Student achievement and understanding are commonly linked with attributes of learning, the latent traits of student learning such as critical thinking, research skills or even the broader concept of content domains in general. Unfortunately, accurately measuring attributes of learning, which, by definition, cannot always be directly measured, has been a consistent problem within the social sciences; requiring assessments to measure the attribute of interest through a set of observable responses (Henson & Douglas, 2005). This difficulty has not hindered interest within measurement, as educational research focused on incorporating cognitive models into test design and analysis has steadily increased, particularly with cognitive diagnosis models (Cui & Leighton, 2009). Beginning with the early models of Fischer's linear logistic trait model (LLTM) in the early 1970s, Tatsuoka's rule space model and Embretson's multicomponent latent trait model (MLTM) in the 1980s, research through to the current day has focused on advancing ways to represent student understanding that can benefit student education and instruction (Cui & Leighton, 2009).

History

Cognitive models originated in the history of computer science as "the simulation of human problem solving and mental task processing" (Leighton & Gierl, 2007, p. 5). Since that origination, cognitive diagnosis models (CDMs) have begun to establish a foundation in educational assessment as the structures of the models are straightforwardly applied in student skills assessment. CDMs used in education have been designed to provide a student with information regarding the mastery, or understanding, of discretely defined skills or attributes which can then be used to identify areas where the student needs revision (Huebner, 2010). For example, Tatsuoka (1993) argued that diagnostic models provided fine-grained analysis of examinee skills and misconceptions that could be harnessed to provide remediation opportunities, particularly useful within formative assessment instances.

Where education harnesses the power of CDMs to identify gaps in student understanding, the implications of CDMs in psychological and psychiatric assessment are somewhat different than education venues, as CDMs can be used to diagnose disorders driven by multiple or combined syndromes, therein providing meaningful information to meet the treatment needs of the patient based on syndromes identified with the assessment (Templin & Henson, 2006). The capabilities inherent to CDMs provide diagnostic elements in psychological measurement, as psychological disorders can be measured as dichotomous attributes, therein allowing the researcher to ascertain the probability of a defined disorder (Templin & Henson, 2006). The information obtained from the attributes has been used to provide diagnostic and structural information simultaneously, aiding in not only diagnosis but the development of disorder manifestation theories as well (Templin & Henson, 2006). The cross-genre application is not without limitations as the assumptions pertaining to the latent variable structure, being compensatory or noncompensatory, are logical within educational assessment; however, these assumptions are not as easily accepted within psychological assessment (Templin & Henson, 2006). The expected functionality of item responses within psychological assessments does not always follow the same structure, therefore, diagnostic application within that environment has to be structured differently (Templin & Henson, 2006). Though valuable in psychology, the remainder of this paper will focus on the use of CDMs in student education.

Future

Diagnostic assessment opportunities continue to increase, particularly within education, as there is a consistent focus on identifying students understanding and, arguably more important, where students are struggling and conversely but complementary, succeeding. The use of large scale assessment has seen a shift in focus within recent years. Stakeholders want feedback that can be used to identify a student's current misunderstanding, implemented to guide modification in educational activities and processes (Gu, 2011). Educational assessments have consistently utilized item response theory (IRT) models to order examinees on a cognitive scale representing levels of ability and therefore provide stakeholders information pertaining to estimates of student ability, whereas CDMs are designed to provide skill or attribute mastery information based on examinee performance to pinpoint strengths and weaknesses in understanding (Junker & Sijtsma, 2001).

By incorporating diagnostic assessment, state education departments, districts and classroom teachers could position themselves to more accurately evaluate and identify student knowledge gaps. Identification of these gaps allows educators to tailor student instruction based on unique or specific needs (Close, 2012). Since the No Child Left Behind (NCLB, 2002) legislation, there has been a requirement for states to provide accountability information across student groups and provide useful assessment information to students, parents, teachers, principals, school districts, state departments of education, etc. in a timely manner (DiBello & Stout, 2007). The signing of the newest education law, Every Student Succeeds Act (ESSA, 2015), continues to require state assessments of student learning, however, there have been some shifts in the key components of NCLB. While ESSA outlines dedicated funding for the lowest performing schools and has developed programs to reward innovation and evidence building, the law allows individual states to determine student performance targets and school rating. As education departments ponder the commitment to Common Core standard alignment, many states are preparing for the evolution of student assessment practices. There is increasing focus on student-level education plans as seen in the Every Student Succeeds Act (ESSA) and the plans delivered by the consortia Smarter Balanced (SBAC) and the Partnership for Assessment of College and Career Readiness (PARCC) that outline the incorporation of diagnostic assessment components within their testing systems (Topol, Olson, Roeber, & Hennon, 2012).

CDMs unique benefit is the models outline student understanding in a different format than customary assessment measures. Educational assessments in the traditional format are designed largely based on unidimensional item response theory and are therefore intended to utilize that response model to create a scale and provide proficiency estimates for students along a continuum (Pellegrino, Baxter, & Glaser, 1999). Fundamentally, unidimensional IRT models provide ability estimates of examinees on a continuous unidimensional latent trait; a higher ability estimate equates to a higher probability of successfully completing the item (Gu, 2011). To provide assessments capable of influencing classroom instruction and individualized student learning, there is a need for test design to expand or shift to incorporate the fundamentals of diagnostic assessment (Pellegrino et al., 1999). As noted within IRT- and Classical Test Theory (CTT) -based assessments, the focus is on obtaining an estimate of ability, however, the goal of CDMs is to provide a measure of student understanding based on mastery of a skill or set of skills (Henson & Douglas, 2005; Huebner, Wang, & Lee, 2009; Xu, Chang, & Douglas, 2003). For CDMs, this information can then be used to clarify the instructional needs of a single student or group of students (Henson & Douglas, 2005). Though the outcomes are different for CDM and IRT models, Tatsuoka (2009) opened the door for cognitive diagnosis models by highlighting the similarities across the models noting that the item response curves of two items representing the same cognitive skill set tend to be similar.

Formative Assessment and Cognitive Diagnostic Assessment

The idea of formative assessment has been in place within education since the introduction of classroom activities and assessments designed to understand a student's current level of understanding. The initial consideration is attributed to Scriven (1967) and was discussed initially within the framework of education evaluation. Not until more extensive treatment by Bloom, Hasting, and Madaus (1983) was assessment concretely defined as being either formative, summative or both. It was years later that Paul Black and Wiliam (1998) published a review of the positive effects of formative *assessment* that propelled formative assessment theory, conversations and application strategies to the forefront of education research. Unfortunately, the increased attention and research over the last two decades has only highlighted the misunderstandings surrounding what formative assessment is and the varied theories regarding its effective application. For this study, the concept of formative assessment was considered in the context of Black and Wiliam's subsequent work meant to define formative assessment within pedagogic initiatives, communication between teachers and students, effective implementation strategies, and attempts to formulate change within classroom instruction (P. Black, 2007; P. Black & Wiliam, 2009; Wiliam & Thompson, 2007). Some years later, Black and Wiliam (2009) revisit their initial definition of formative assessment, while applying terminology presented by the Assessment Reform Group (ARG, 2002), to establish an understanding of formative assessment as:

> Practice in a classroom is formative to the extent that evidence about student achievement is elicited, interpreted, and used by teachers, learners, or their peers,

> to make decisions about the next steps in instruction that are likely to be better, or better founded, than the decisions they would have taken in the absence of the evidence that was elicited. (p.7)

The authors provide an additional, clarifying statement on formative assessment as:

> "…it is clear that formative assessment is concerned with the creation of, and capitalization upon, 'moments of contingency' in instruction for the purpose of the regulation of learning processes. This might seem to be a very narrow focus, but it helps to distinguish a theory of formative assessment from an overall theory of teaching and learning." (p.8)

Interestingly, the common beliefs about the goals of formative assessment lie in research completed many years prior to Black and Wiliam's 1998 work which is the modern day most cited work within formative assessment scholarship. Prior to their review, work published regarding the processes or practices necessary in teaching to support learning centered around establishing where the learners are in their learning, where they are going and what needs to be done to get there (Ramaprasad, 1983). These three processes are the key elements describing the goal(s) of formative assessment today. Though written and defined independent of the concept, Wiliam and Thompson (2007) applied these three processes to identify and support five key strategies of formative assessment by their role within education; including the teacher, peer and learner (see Table 1). The five strategies include; sharing success criteria, classroom questioning, comment-only marking, peer- and self-assessment, and the formative use of summative assessment (Wiliam & Thompson, 2007). The three processes considered with these five strategies outline how formative feedback can provide the needed information to formulate change and promote learning. Early work within formative assessment outlined several classroom activities developed with and by teachers which were found to be effective and included: making sure students knew the criteria(s) for success, activities that promoted classroom questioning, comment-only marking, peer- and self-assessment, and finally incorporating summative tests for formative purposes (Black and Wiliam, 2009).

Table 1. Wiliam and Thompson's 2007 formative assessment strategies framework

	Where the learner is going	Where the learner is right now	How to get there
Teacher	Clarifying learning intentions and criteria for success	Engineering effective classroom discussions and other learning tasks that elicit evidence of student understanding	Providing feedback that moves learners forward
Peer	Understanding and sharing learning intentions and criteria for success	Activating students as instructional resources for one another	
Learner	Understanding learning intentions and criteria for success	Activating students as the owners of their own learning	

The differentiating component of formative assessment is based on the premise that formative assessment practices occur during instruction and can afford educational adjustments during the learning process. The adjustments are made during student learning, and may include "real-time" modifications made during one-on-one teaching, within small group activities, or during classroom discussions and activities (Black and Wiliam, 2009). Teacher feedback provided through grading practices or evidence from student work can influence decisions made in planning subsequent lessons, as can lessons learned from activities taught in previous years (Black and Wiliam, 2009). The purpose of the adjustments are to promote and encourage student thought processes while improving and developing student learning, resulting in an increase in student understanding and achievement. As the goal of formative assessment in the end is to improve student learning, a key component of that goal is a student understanding of how they are progressing (Sadler, 1989). This feedback is crucial for a student to be aware of and understand the criteria and standards for success, including information about quality of performance, what the student understands as well as identifying areas needing remediation.

Feedback to the student is an essential component of formative assessment that promotes student understanding, learning and achievement (Hattie and Timperley, 2007). In their 2006 study, Black and Wiliam noted, “the quality of interactive feedback is a critical feature in determining the quality of learning activity, and is therefore a central feature of pedagogy” (p.100). Feedback in education is most effective when it identifies what a student currently understands and what still needs to be understood, therein, highlighting the process of learning that is currently taking place; and, defining the student’s current location in relationship to the end goal and providing information about the performance on a task and how it can be completed more successfully (Sadler, 1989; Hattie and Timperley, 2007). This again aligns with the three key processes presented by Ramaprased (1983). Like Wiliam and Thompson (2007), Hattie and Timperley (2007) incorporated Ramaprased’s three processes, but as three questions that must be answered for feedback to be influential and effective: Where am I going? How am I doing?, and Where to next? By answering these three questions, effective feedback clarifies to the student what the immediate goals are, what progress has been made toward the goal(s), and what needs to happen to make better progress toward the goal (Hattie and Timperley).

Considering these efficiencies in the context of assessment feedback has been problematic, particularly in terms of identifying the effectiveness this type of feedback has on student understanding. The power of an assessment has typically been viewed as the end goal, or *summative*, of learning, the measurement that provides information about what the student should be able to successfully do after a given period of learning has taken place (Hattie and

Timperley, 2007). The challenge with assessment feedback is that though there is information outlining where a student is going (what the goals are), there is rarely direct language linking where a student is and what are the needed next steps (Hattie and Timperley). Most assessments, formative or summative, are typically not developed, offered, or scored using statistical models capable of providing the needed feedback about the skills a student has mastered and not mastered to outline where a student is and what needs to happen to reach the end goal; it is this gap that CDMs could fill. And, it is that gap that this study addresses.

CDMs are designed primarily to identify information about why an examinee is not succeeding in content areas using defined skills or attributes and an item-to-attribute alignment matrix known as a Q-matrix. Educational assessments designed using a CDM utilize cognitive theory relying on statistical modeling to provide inferences about the student mastery of the attributes measured by the items (de la Torre, 2008; Jang, 2009). By accurately and explicitly defining the measured attributes and creating a Q-matrix, a test designed to be diagnostic pinpoints each student's mastery level by evaluating performance through the item response probabilities linked to the attributes or skills being measured (de la Torre; Jang). von Davier (2005b) described the goal of CDMs:

> "….is to identify skill profiles, that is, to perform multiple classifications of examinees based on their observed response patterns with respect to features (skills/attributes) that are assumed to drive the probability of correct responses." (p.1)

To provide this level of information, the design of the diagnostic assessment is in contrast to traditional assessments that are designed using item response theory (IRT) and classical test theory (CTT) that provide a measure of ability that summarizes the examinee's ability with a single defined latent trait (Henson, Templin, & Willse, 2009; Xu et al., 2003). As the two main features of CDMs are (1) the attributes being measured by the assessment and (2) the Q-matrix that identifies which items measure each attribute, most summative educational assessments contain little diagnostic information as they are designed using the conventional unidimensional framework of IRT or CTT resulting in scaled scores that represent the ordering of students along a unidimensional continuum (Henson & Douglas, 2005). As Henson et al. (2009) noted, extracting diagnostic information (mastery/non-mastery) from such designed tests would require additional analyses. Though diagnostic assessments provide feedback that is different from IRT-based results, it is important to note many CDMs are essentially extensions of IRT models designed to classify examinees according to latent characteristics (Rupp & Templin, 2008a). CDMs have been defined as special cases of latent class models, known as multiple classification latent class models, and have been used to characterize the relationship of item responses to a set of attributes using the item to attribute alignment defined by the Q-matrix (Templin & Henson, 2006).

Attributes

As noted, diagnostic assessments combine cognitive theory with statistical models to make inferences about test takers' mastery of tested skills, also referred to as attributes (de la Torre & Karelitz, 2009; Jang, 2009). In education, attributes can be thought of in more general terms as representations of knowledge or understanding, as cognitive processes, or knowledge states (de la Torre, Hong, & Deng, 2010). For example, in the context of mathematics, an attribute would be the ability to divide four-digit number or graphing basic exponential functions. Attributes differ from the scaled scores of general ability measurement, used with most large-scale assessments as attributes are categorical latent variables that can be classified to represent either master or non-master of the attribute (de la Torre & Karelitz; Templin, Henson, Templin, & Roussos, 2008). The term ‘attribute or skill profile’ is the diagnostic feedback characteristic of the diagnostic assessment, used to identify an individual student’s mastery of tested skills, these profiles are represented as binary values to indicate skill mastery or non-mastery (de la Torre & Karelitz; Jang).

Within education, the value of cognitive diagnostic assessment and attribute profile is clear. One example was described by Henson and Douglas (2005), wherein a classroom teacher was interested in evaluating student attribute profiles, through the use of a set of dichotomous attributes that outlined student mastery and non-mastery for each attribute. By looking at the individual student information in the attribute profiles for each student, the teacher was also able to identify instructional needs for the entire class (Henson & Douglas).

Q-Matrix

As noted, a key component of CDMs is the Q-matrix which is the attribute-by-item matrix that defines how each item is associated to each attribute; a hypothesized linking of items and attributes (de la Torre et al., 2010; DiBello, Roussos, & Stout, 2007). Similar to blueprints in summative assessment development, the Q-matrix embodies the design of the assessment, serving as the attribute blueprint for assessment construction in cognitive diagnostic assessments; defining (by item) which attribute(s) are required mastery to increase probability of a correct response (Close, 2012; de la Torre et al.; Henson et al., 2009; Liu, Xu, & Ying, 2010). For the structure of a Q-matrix, each item is a row and each attribute, or skill, is a column; if an item measures a skill, there is a one in the corresponding item × attribute cell, if it does not, the cell contains a zero (Close; de la Torre et al.; Henson & Douglas, 2005; Henson & Templin, 2007; Henson et al.). The completed matrix provides the “map” of assessment items to measured skills, not unlike test blueprint for non-diagnostic assessments. However, whereas summative assessment blueprints indicate what skills are being measured by the item, a Q-matrix defines what skills must be mastered in order to answer the item correctly. CDMs are structured to provide the foundation for calculating the probability of a correct response values by pairing the Q-matrix information with the attribute profile (Henson et al.).

Model Compensation

Model compensation is a differentiating factor between CDMs, as models differ in the dependency of attributes when responding to items (de la Torre & Karelitz, 2009; Henson et al., 2009). Model compensation describes this dependency and the process by which skills are applied when answering items (Henson et al.). Within model compensation there are two over-arching models: compensatory and non-compensatory. Compensatory models do not have a conditional dependency on attribute mastery, meaning, compensatory models allow an individual to “compensate” for non-mastery of one skill by having mastered another (Henson et al.). Noncompensatory models can be viewed in the opposite as the models require mastery of all attributes measured by the item and can include conjunctive and disjunctive models (Henson et al.). Because the conditional dependency is not present in compensatory models, the log-odds of a correct response remains constant across all levels of mastery for the other required attribute(s) measured by the item (Henson et al.).

The conjunctive and disjunctive requirements within noncompensatory models are similar to model compensation. Conjunctive models require students to successfully use all attributes measured by an item to answer correctly while a disjunctive model merely requires competency on any one of the measured attributes for probability of answering correctly to be high (Close, 2012; DiBello et al., 2007). In a conjunctive model, if a student lacks competency on any measured attribute for an item, the probability of successfully completing the item is lower as mastery of one attribute cannot counteract non-mastery of other attributes; the student must be a master of all required attributes (Close; DiBello et al.; Henson et al., 2009) On the other hand, in disjunctive models, the interaction of skill and responses is modeled to create a high probability of a correct response when at least one sufficient skill is mastered (Close; Henson et al.). For example, an item measures two skills; with a conjunctive model applied, a student who has mastered both skills will have a higher probability of correctly completing the item than a student who has only mastered one of the two measured skills. Using this same example under a disjunctive application, the probability of successfully completing an item increases with the mastery of just one of the two measured skills. This can be expressed as an item can be answered successfully using different strategies and answering the item correctly only requires that one of the strategies is used (DiBello et al.). The difference lies within the model's calculating the probability estimation of correctly answering the item; for conjunctive, if two skills are measured, two skills must be mastered to result in an increase in the probability estimation, for the disjunctive, mastering one of the two measured skills would results in an increase in the probability of correctly answering the item.

As states finalize testing program activities from the current school year, it is apparent several things within educational assessment are on the verge of dramatic change. Educational news sources, social media, professional blogs and legislative dockets outline the perceived failures of consortia promises, misconceptions, and impatience of parents and tax payers while highlighting the demand for change. States are abandoning their commitment to participate in the prescribed program of educational consortia, are teaching new standards, writing items designed to measure student ability more accurately, incorporating technology throughout learning and assessment, and trying to disseminate student results in a manner that can provide explicit information regarding student understanding and influence current learning activities. This has reinforced the plea for assessment data that can provide information that is more aligned with criterion-referenced tests; providing feedback about student learning that is capable of supporting individualized, tailored education through identification of what a student understands and, more importantly, where there is confusion and struggle.

This demand for student data shifts from ex post facto assessment to times during learning, through assessment that outline performance on specific components of learning or cognitive areas (Henson et al., 2009). Cognitive diagnostic assessments are designed to provide this information by establishing the relationship between student response and attribute mastery; measuring the knowledge structures and the processing skills to provide information about the cognitive strengths and weaknesses of students (J. Leighton & M. Gierl, 2007). Diagnostic models promise results that can be used to classify students based on skill-level performance as opposed to where the student lies on a continuous distribution (Rupp & Templin, 2008a). CDMs also deliver precise information about student comprehension that can be used to formulate classroom activities, construct individualized student learning plans and identify areas of misunderstanding at the student or group level for remediation purposes when it is most valuable (Close, 2012; Cui & Leighton; de la Torre & Karelitz, 2009; Henson et al.).

Statement of the Problem

The purpose of this study is to utilize a formative assessment data set to evaluate and validate the hierarchical diagnostic classification model (HDCM) as a tool in identifying learning hierarchies. Because learning hierarchies are potentially present in several educational and psychological scenarios, this research is valid as fitting assessment data to models that do not account for the hierarchical structure causes the model to over fit the data. This research will assess the effectiveness of HDCM as a “middle-man analysis” wherein data will be evaluated using the Log-linear Cognitive Diagnostic Model (LCDM), modified to account for hierarchies and evaluated using the HDCM to evaluate improvement in model fit and classification accuracy; therein highlighting the potential benefit of HDCM to serve as a statistical way to test for an attribute hierarchy.

The rationale for this dissertation is grounded in an extensive and in-depth review of the literature and intended uses of cognitive diagnostic assessments and models. There is reason to hypothesize that some learning processes occur linearly, requiring the learner to master one concept before progressing to the next. In addition, the current research dictates that there is need for statistical models capable of testing for the presence of learning attribute hierarchies that can then be used to accurately and appropriately identify attribute hierarchies for use with models such as the Rule Space model or the Attribute Hierarchy Method. Moreover, there is significance of this study in extending the HDCM work done by Templin and Bradshaw (2013) from language acquisition to K-12 assessment practices.

Hypothesis

There is an attribute hierarchy present in how students learn and understand the essential concepts of slope; consequently, the HDCM will fit the data from a formative mathematics assessment designed to measure the hierarchy better than the LCDM.

Research Questions

1. Using a formative assessment designed to measure a specific and well-defined learning hierarchy, does the HDCM accurately identify the presence or absence of the hierarchy?
2. What are the estimations and classification differences between the HDCM, LCDM and AHM from a formative assessment designed using AHM to measure a learning hierarchy?

CHAPTER 2

Literature Review

Cognitive diagnosis assessments are designed to provide inferences about an examinee's level of mastery within a well-defined skill set and provide the user with a "skill profile" outlining the examinee's mastered competencies, misconceptions and erroneous strategies within that skill set (Gu, 2011; Jang, 2009). It is this individualized student feedback capability that draws users to cognitive diagnosis models (CDMs) and sets these assessment models apart from other assessment designs, particularly within formative assessment. Driven by the potential to inform feedback and increases in understanding, CDMs are emerging as effective measurement and classification tools in education, psychology, industry, and health sciences (Rupp et al.). Assessments designed using diagnostic models promise to measure an examinee's ability, or skill mastery, of a predetermined latent variable or set of latent variables and yield examinee classification(s) based on the response pattern (Rupp, Templin, & Henson, 2010). Because the assessments involve a clearly latent identified variable or set of variables, CDMs have the capability of measuring fine-grain student abilities, while providing feedback to users about current knowledge, performance and ability that can be used to foster unique student learning environments focused on results. The application potential is vast as CDMs have been meaningfully applied to provide clinical and neurological diagnosis, cognitive diagnosis in education, as well as standards-based assessment practices in education (Rupp et al.).

Cognitive Diagnostic Assessment

Introduction and Overview

Cognitive diagnostic assessment is defined under several structure labels: cognitive diagnostic models, cognitive psychometric models, and cognitively diagnostic models (Rupp &

Templin, 2008b). Additionally, Rupp and Templin noted the models can be considered extensions of item response theory models (Embretson & Reise, 2000) and have been referenced within that research base as cognitive diagnosis models (Nichols, Chipman, & Brennan, 2012), multiple classification models (Maris, 1999), restricted latent class models (Haertel, 2005), and structured item response theory models (Rupp & Mislevy, 2007). Though the terminology presented to label diagnostic assessment may vary, the intended functionality remains consistent. The principal goal of CDMs is to deliver several items intended to assess a specific skill set supported by a matrix that defines the relationship between the items and the skills necessary to successfully answer the items (von Davier, 2005a). This matrix is identified as the Q-matrix and is a fundamental component of all diagnostic assessments as it provides the intended attribute structure being tested (Rupp et al., 2010). This structure is designed to identify cognitive strengths and weaknesses by evaluating examinee processing skills and knowledge structures through test items (Leighton & Gierl, 2007). By linking response probabilities with specified skills, CDMs provide diagnostic feedback about the examinee's mastery level for the latent variables identified for the particular assessment (Jang, 2009).

The aim of CDMs can be simplified to the classification of an examinee, based on observed examinee response patterns in relationship to the measured attributes, to provide more concise information about the examinee's current level of understanding (von Davier, 2005a). This can be generalized as Leighton and Gierl (2007) defined cognitive models in educational assessment as the "simplified description of human problem solving on standardized educational tasks, which helps to characterize the knowledge and skills students at different levels of learning have acquired and to facilitate the explanation and prediction of students' performance" (p. 6). More specifically, CDMs are designed to utilize proficiency assessment tools to ascertain the presence or absence of explicit skill sets and can identify a student's current understanding and level of learning (de la Torre & Karelitz, 2009). This is accomplished as CDMs are a special class of latent models wherein an examinee's ability estimate is used to model the probability of correctly answering an item (Henson & Douglas, 2005; Henson, Roussos, Douglas, & He, 2008). This estimation lends itself well to education and the current desire to acquire finite information about student success and confusion. The latent variables within CDMs are predominantly categorical, often referred to as skills or attributes, and are typically dichotomous (Templin et al., 2008). This structure allows the examinee to be classified as master or non-master by evaluating the relationship between observed data and the set of latent variables (Templin & Henson, 2006; Templin et al.).

Design Specifications

Cognitive diagnosis models can provide users with not only diagnostic feedback about an examinee's strengths and weaknesses, but with detailed information about the processes examinees pass through as they complete each item (Tatsuoka, 1993; Zhou, Gierl, & Cui, 2009). Essentially, a well-designed CDM can provide a user with precise information about an examinee's problem-solving and mastery of a specified set of skills and can be integrated with instructional methods and learning processes to improve mastery (Tatsuoka, 1993; Zhou et al.). The results provide additional information for educators to ascertain students' current level of understanding and assist in remediation and instructional activities (Tatsuoka). This information can then be used to create student education plans drafted to focus efforts on skills a student is struggling to succeed with.

The complexity of CDMs is a caveat for implementing the models, as extensive knowledge and expertise is required regardless of model as several elements need to be considered when developing an assessment for cognitive diagnosis. Initially the goal of the assessment and the criteria for diagnosis need to be considered, tasks need to be designed or selected to gather information about an examinee's competencies based on the diagnosis criteria, a scoring and reporting system must be in place to accurately provide diagnostic information at the attribute level and the usefulness of the reporting must be considered (Jang, 2009). To provide accurate information this process begins by evaluating and studying the processes and strategies examinees must use to respond to an item (Zhou et al.). Given appropriate decisions, there are several models that can be implemented based on evaluation needs.

Diagnostic Models

Research that is designed to result in examinee classification, driven by responses according to multiple latent class variables, can be completed with one of many diagnostic models (Rupp & Templin, 2008b). These models envelop three defining characteristics including the structure of the response and latent predictor variables (dichotomous vs. polytomous), as well as the compensation rules associated with the latent predictor variables for item performance (Rupp & Templin, 2008b). Based on these characteristics there are several models frequently used within cognitive diagnostic assessment research that can be partitioned as compensatory, non-compensatory and general. Original models include Tatsuoka's rule space method (Tatsuoka, 1985) and the Attribute Hierarchy method (AHM) (Leighton, Gierl, & Hunka, 2004). The oft-cited non-compensatory models include the deterministic-input, noisy-and-gate (DINA) model (de la Torre & Douglas, 2004; Haertel, 2005; Junker & Sijtsma, 2001; Rupp & Templin, 2008a), the noisy-input, deterministic-and-gate (NIDA) model (Junker & Sijtsma, 2001), the non-compensatory reparameterized unified model (C-RUM) or Fusion model (DiBello, Roussos, & Stout, 2007; Hartz, 2002), the reduced NC-RUM (Templin, 2006), the higher-order DINA

(HO-DINA) (de la Torre & Douglas), and the multi-strategy DINA (MS-DINA) model (de la Torre & Douglas).

The compensatory models include the noisy-input, deterministic-or-gate (NIDO) model (Templin, 2006), the deterministic-input, noisy-or-gate (DINO) model (Templin, 2006; Templin & Henson, 2006), and the compensatory reparameterized unified model (C-RUM) (Hartz, 2002; Templin & Henson, 2006). The more general models, which allow the user to organize and estimate compensatory and non-compensatory CDMs, include the log-linear diagnostic model (LCDM) (Henson et al., 2009) and the general diagnostic model (GDM) (von Davier, 2005a). Of these, the three most commonly discussed and implemented are the DINA, the LCDM and the GDM (von Davier). The multitude of options requires the researcher to evaluate the intent or purpose of the assessment, the structure and design, as well as the types of constraints that will be placed on the model. These constraints will drive the selection of the model as each model specifies the structure of the constraints to be implemented (Templin et al., 2008). Rupp and Templin (2008b) provided nine defining characteristics of DCM to assist researchers in evaluating and selecting appropriate models, including: the multidimensional nature, the confirmatory nature, complexity of the loading structure, type of response variable suitable, contained latent predictor variables, nature of latent predictor variable interactions, allowable criterion-references interpretations, and types of heterogeneity modeled.

Q-Matrix

The key premise of diagnostic testing is that every item within a test evaluates an attribute or skill, or set of attributes/skills and can be represented in an item-skill relationship matrix (von Davier, 2005a). This matrix, commonly referred to as a Q-matrix, is the foundation of many diagnostic models; defining each attribute pattern needed to successfully complete an item (von Davier). The matrix essentially functions as the assessment design, highlighting which attributes are required per item (Close, 2012).

Within diagnostic models, the Q-matrix outlines the relationship between the attributes and the items, identifying which attributes are measured by each item (Rupp et al., 2010). Within this incidence matrix, each attribute is a row within the matrix, while the items are represented by columns wherein the attribute(s) for that item are noted (Gierl, 2007). The Q-matrix is the essential characteristic of diagnostic assessments, as proper item/attribute identification is critical for effective diagnosis (Close, 2012; Rupp & Templin, 2008a). Across models, the matrix identifies attribute (*k*) by item (*j*) requirements, wherein *qjk* highlights if mastery of the *k*th attribute is required for the *j*th item (Henson & Douglas, 2005; Henson et al., 2009).

$$qjk = \begin{cases} 1 \text{ if item } j \text{ requires attribute } k \\ 0 \text{ if else} \end{cases}$$

Each model identifies the functionality of the Q-matrix in relationship to the additional estimation parameters based on compensation rules of the model. For example, within DINA, where mastery of all attributes associated within an item is required to correctly answer that item, each student is classified as having mastered all of the required attributes or not, represented as a vector of indicators, *αi* (Henson & Douglas, 2005). In the model, each indicator within the 0/1 *αi* vector is represented as *ξij*, noting mastery of all required attributes for item *i* (Henson & Douglas).

$$\xi ij = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}}$$

Consequently, estimating the probability of a correct response only requires two parameters, slipping and guessing (Henson & Douglas). These parameters are estimated based on model specifications.

Model Compensation

A major directive decision when choosing a model involves the compensation structure of the model. As noted previously, there are several models that fall within each of the compensation structures. Model compensation refers to the manner in which the model allows skills an examinee possess to influence probability of success patterns (Gu, 2011). Compensatory models combine the latent predictor variables to allow the presence of one attribute to "compensate" for the lack of another (Rupp & Templin, 2008b). Conversely, non-compensatory models do not provide adjustments to the probability of success based on attribute mastery. As noted in the definition, non-compensatory models require each specified skill in the set be present for successful item completion (Rupp & Templin, 2008b).

In compensatory models, the interaction of a required attribute and the item is not dependent on any additional attribute mastery (Henson et al., 2009). Consequently, the probability of a correct response remains constant across mastery levels of the other required attributes, meaning the likelihood an examinee correctly answering an item does not increase simply because they have mastered more than one of the required attributes (Henson et al., 2009). For example, in the Compensatory RUM (Hartz, 2002) the lowest probability of a correct response, similar to the guessing parameter in IRT models, is defined as $–\pi^*_j$ (Henson et al.). This probability of a correct response, defined as r^*_{jk}, increases as a function of each required attribute that is mastered, resulting in a relationship of item performance and required attribute that is not conditional on any other required attributes (Henson et al.).

Non-compensatory models assume the lack of a required skill influences the probability of a correct response regardless of other Q-matrix identified skills an examinee does possess (Gu, 2011). As the name suggests, these models do not allow for deficits in one skill to be "compensated" for by another skill (Rupp et al., 2010). Consequently, noncompensatory models do not make adjustments in the probability estimation based on additional skill mastery. Several models are consistently used and are clear in their compensation rules and perspective level. The DINA and NIDA models are both non-compensatory, however, DINA is a more complex model as it is designed at the item-level perspective, whereas NIDA functions at the skill-level perspective (Kim, 2011). This difference in complexity is driven by the fact that there are always more items than there are skills to be measured (Kim).

Model Condensation

Quite similar to model compensation, model condensation is another component used to drive model choice. Within diagnostic testing, typically items within a test are designed to evaluate mastery of at least one skill, generally multiple skills are evaluated. When an item combines multiple latent response variables, meaning an examinee would need multiple skills to correctly answer the item, the variable combination is *condensed* to result in a single response (Rupp & Templin, 2008b). In noncompensatory models there is a conditional relationship between the attribute and the item responses that, by model design, is dependent on other attributes; therefore, these models can be defined through condensation as well (Henson et al., 2009). This condensation defines how the model compensates for attribute mastery and provides rules for the products of the latent variables, most often as conjunctive and disjunctive (Rupp & Templin, 2008b).

Disjunctive models can be defined as models that associate the mastery of at least one skill, generally a subset of skills, with a high probability of a correct examinee response on an item (Close, 2012; Henson et al., 2009). In conjunctive models, probability of a correct response is directly associated with mastery of all required skills for a given item and lack of mastery for any one of the required skills greatly reduces the probability of a correct response (Close). In other words, when an item requires a specific set of attributes, lacking any one of those attributes cannot be made up for, or compensated for, by mastery of another required (Henson et al.; Jang, 2009). Within conjunctive models, correctly solving an item is reliant on successful completion of a series of steps, for example, correctly completing a mathematical item requires completing a collection of processes or skills to answer the item correctly (Kim, 2011; Roussos et al., 2008).

Rule Space

As the desired output of assessment practices began to converge with cognitive theory, there was a need to enhance assessment practices in a manner that could provide meaningful information regarding the knowledge, processes and strategies applied to correctly answer test items (Embretson, 1983). A concept which was later reiterated as the cognitive design system (CDS) was developed to outline the process and framework necessary to establish the link of cognitive theory to test design and examinee performance needed for construct validation (Embretson, 1995). In developing the CDS, Embretson established a three-stage process that has been used to validate several constructs in language, reading and mathematics (Leighton et al.). The steps progress from describing the goals of measurement to establishing the construct representation and conducting nomothetic span research; wherein construct representation refers to breaking down a task in a way that allows the researcher to identify the processes, strategies and knowledge needed to respond to an item, while nomothetic span research seeks to identify the relationships between the test score and other measures (Embretson, 1983). As Tatsuoka's rule space method was the initial cognitive diagnosis model, clarifying the functionality of the model is appropriate to outline the groundwork for subsequent models. Tatsuoka's rule space method can best be thought of as a system rather than a single model in that the method provides skill diagnosis information as well as offering knowledge improvement (Kim, 2011). This model classifies student responses as sets of mastery and non-mastery patterns (Gierl, 2007). The model was designed to support and highlight information pertaining to an examinee's mastery of a specified set of cognitive skills by outlining an organized procedure to measure attributes, identifying the diagnostic performance of the skills on an assessment for test developers, and providing cognitively-based results intended to enhance student learning and instruction (Gierl). The model is driven by two elements: the Q-matrix and rule space, which is accomplished by examining the response patterns within a geometric space (Kim).

As noted previously, the Q-matrix is a binary matrix that outlines the item-by-skill and indicates which skills are associated with each item (Tatsuoka, 1985). By representing skill mastery in a binary vector of latent variables, the user is able to ascertain the presence or absence of each skill within the set under diagnosis (de la Torre & Douglas, 2004). By evaluating the response patterns, which are representative of the examinee's current ability and knowledge state, it is possible to assess misunderstandings, which are represented as ill-fitting response patterns (Tatsuoka, 1985). The rule space method looks to identify atypical examinee responses in a Euclidean space, identifying mastery essentially through error analysis (Close, 2012). Using the two-dimensional space of theta (ability) and zeta (unusual response patterns identified as aberrant) which are used to create a "rule space" that identifies the ideal response pattern (Gierl, 2007). Utilizing Mahalanobis distance, rule space is represented by measuring the distance between ideal and observed response patterns (based on theta and zeta values), wherein the observed patterns lying closest to the ideal represent student understanding (Gierl; Kim). By identifying the misconceptions, or the observed patterns farthest from the ideal patterns, the feedback can provide meaningful information for remediation plans (Tatsuoka, 1983). Essentially, the rule space method essentially creates “ideal” patterns for which the examinee’s ability is then mapped to the nearest “ideal” response pattern; as such, this model serves as a statistical pattern recognition method (Templin & Bradshaw, 2013).

The rule space model does have its limitations, understanding the extensive number of possible response patterns and the variability associated with those patterns, there is reason to question the accuracy of identifying the examinee’s knowledge state (Kim, 2011). Though several models advanced from Tatsuoka’s Rule Space model, the one to incorporate the assumption of a learning hierarchy was the Attribute Hierarchy Model.

Attribute Hierarchy Model

As cognitive research advanced, popular theory suggested that cognitive skills function through a network of interrelated processes, suggesting and assumption of attribute dependence, consequently, a need arose to formulate an extension to Tatsuoko’s rule space model (Leighton, Gierl, Hunka, 2004). Attribute Hierarchy Method (AHM) provides that extension by incorporating the assumption of an attribute hierarchy using a cognitive item response theory (IRT) model (Leighton et al). A cognitive IRT model provides the desired link between cognitive and psychometric models through parameters that provide information specific to the demands of the items and a student’s ability level (Leighton et al.). As evident by being defined as an extension, AHM utilizes the observed response patterns to classify students against ideal response patterns and incorporates the matrices applied in rule-space research, including the adjacency, reachability, incidence and reduced Q) to produce the needed ideal response patterns; both of which are signature components of the rule-space model (Leighton et al.).

To correctly apply the AHM, the attribute hierarchy must be identified prior to test development and would be used to guide item development (Leighton et al., 2004). Under rule-space, the researcher typically identifies attributes for items that have already been created using the incidence matrix, which identifies the the attributes involved in completing each item (Birenbaum and Tatsuoko, 1993). The cognitive patterns represented within an incidence matrix provide the ideal item-score patterns or knowledge states that are then used to match with observable item-score patterns (Birenbaum and Tatsuoko).

Once identified, the researcher is able to create an adjacency matrix to represent the direct relationships among the measured attributes of order (k,k) where k is the number of attributes (Leighton et al., 2004). The matrix serves to identify the attributes that are considered to be prerequisites of another attribute as theorized by the attribute hierarchy. Each row and column of the matrix is an attribute, with a one in the row indicating that the row attribute is a prerequisite of the column attribute. For example, Leighton, Gierl and Hunka (2004) identified a six attribute hierarchy wherein attribute 2 and 4 were dependent on attribute 1, attribute 2 a prerequisite for attribute 3 and attribute 4 was theorized to be a prerequisite for attribute 5 and 6 (see Figure 1).

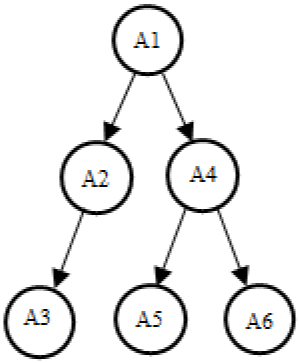


Figure 1. Divergent attribute hierarchy from Leighton, Gierl, and Hunka 2004 showing a theorized attribute hierarchy where A2 and A4 are dependent on A1, A3 is dependent on A2 and A5 and A6 are dependent on A4.

The adjacency matrix to present the attribute hierarchy shown in Figure 1 is below (Matrix 1). For an adjacency matrix, a 1 in the position (j,k) indicates that attribute j is directly connected to attribute k as a prerequisite (Leighton et al., 2004). For example, reviewing the theorized hierarchy in Figure 1, the values in the first row (j) of the matrix represent the dependencies of A1, consequently, there is a 1 in the second and fourth positions (k) to identify A1 as a prerequisite to A2 and A4.

Matrix 1. Adjacency Matrix

	A1	A2	A3	A4	A5	A6
A1	0	1	0	1	0	0
A2	0	0	1	0	0	0
A3	0	0	0	0	0	0
A4	0	0	0	0	1	1
A5	0	0	0	0	0	0
A6	0	0	0	0	0	0

Once the adjacency matrix is completed, a reachability matrix (*R*) of order (k,k) is developed to identify the direct and indirect relationships between the attributes and used to create a subset of items conditioned on the attribute hierarchy (Leighton et al., 2004). This matrix is calculated using either a series of Boolean additions or by using $R = (A + I)^n$, with A being the adjacency matrix, I the identity matrix and n indicating the number required for R to reach invariance and can represent the numbers 1 through k (Leighton et al.). Each row of the matrix identifies all of the attributes, including the attribute of the row, for which that attribute is a direct and indirect prerequisite (Leighton et al.). For example, the first row of Matrix 2 identifies A1 as a prerequisite to all attributes in the hierarchy, whereas the fourth row identifies A 4 as a prerequisite to A4 as well as A5 and A6.

Matrix 2. Reachability Matrix (*R*)

$$
\begin{array}{ccccccc}
 & \text{A1} & \text{A2} & \text{A3} & \text{A4} & \text{A5} & \text{A6} \\
\text{A1} & 1 & 1 & 1 & 1 & 1 & 1 \\
\text{A2} & 0 & 1 & 1 & 0 & 0 & 0 \\
\text{A3} & 0 & 0 & 1 & 0 & 0 & 0 \\
\text{A4} & 0 & 0 & 0 & 1 & 1 & 1 \\
\text{A5} & 0 & 0 & 0 & 0 & 1 & 0 \\
\text{A6} & 0 & 0 & 0 & 0 & 0 & 1
\end{array}
$$

A researcher applying the AHM would now need to create an incidence matrix or Q matrix to identify the potential set of unique items using $2^k – 1$ to determine the number of items needed (Leighton et al.) The Q-matrix serves as a listing of the number of items needed to measure all possible attribute combinations, outlining each item by identifying the attributes that are required to correctly complete the item (Leighton et al.). Continuing with the previous example, as there are six attributes, the Q-matrix for this example would be a 6 by 63 matrix as there are six attributes and $2^6 – 1 = 63$. Each row of the matrix represents an attribute, but more importantly, each column represents a unique item and the values within the column identify the attributes an examinee must possess to successfully complete the item. For example, the third column or item of the matrix indicates that the item would measure attributes 1 and 2, while the 63 column indicates the item would measure all of the attributes.

Matrix 3. Incidence (*Q*) Matrix

```
101010101010101010101010101010101010101010101010101010101010101
011001100110011001100110011001100110011001100110011001100110011
000111100001111000011110000111100001111000011110000111100001111
000000011111111000000001111111100000000111111110000000011111111
000000000000000111111111111111100000000000000001111111111111111
000000000000000000000000000000011111111111111111111111111111111
```

When there is a proposed attribute hierarchy identified, some of the potential items would not be appropriate based on the structure of the hierarchy. For example, the binary values for the fifth item in the Q-matrix is (101000), indicating the proposed item would require an examinee to have mastered A1 and A3, however, the hierarchy identified A3 as dependent on A2, therefore, the item would need to be identified as (111000), which is a duplication of item seven in the matrix and would consequently need to be removed by creating a reduced Q matrix (Q_r) (see Matrix 4). Applying the dependencies specified by the hierarchy, each item the must be re-identified and shown to be duplicative would need to be removed in the reduced matrix. By evaluating the items within the incidence matrix and removing duplication and items that do not follow the identified dependencies of the hierarchy, the researcher is left with the greatly reduced matrix (Q_r) that now serves as the blueprint of the assessment during development (Leighton et al.).

Matrix 4. Reduced (Q_r) Matrix

111111111111111
011011011011011
001001001001001
000111111111111
000000111000111
000000000111111

The Q_r now represents the minimum number of items that must be created to follow the attribute dependencies outlined by the hierarchy (Leighton et al.). In this example, the 63 items included for the incidence (Q) matrix has been reduced to a 15-item matrix, establishing the threshold number of items needed for the hierarchy. For example, the fifth column of the matrix has a binary value (110100) that indicates there needs to be an item developed that measures A1, A2 and A3. This is appropriate given the hierarchy stipulates that A3 is dependent on A2 which is in turn dependent on A1, consequently, this matrix has no redundant or attribute combinations outside of the structure identified in the hierarchy. This matrix establishes the cognitive requirements identified in the attribute hierarchy and the minimum number of items needed to address the possible attribute combination therein; in doing so, AHM incorporates cognitive theory during the test design and development process (Leighton et al.).

The AHM process addresses the examinee response patterns slightly different than Tatsuoka's model. As rule-space theory identifies response patterns as "ideal" examinee response patterns, the shift in terminology to "expected" examinee response patterns addresses the awareness within AHM that the response patterns would be observed if the specified hierarchy identified in the adjacency matrix is true (Leghton et al., 2004). In addition to expected response patterns there are expected examinees, who exhibit attributes congruent with the attribute hierarchy (Leighton et al.).

Continuing with the example from Leighton, Gierl and Hunka (2004), the expected examinee response patterns, total number correct for the assessment, and the examinee response attributes are displayed in Table 2. Within the table, each row represents an expected examinee and identifies the attribute(s) said examinee has (expected attributes), followed by the expected response pattern of that examinee based on the what attributes the examinee has, in turn leading to the number of items the examinee is expected to answer correctly (total score) (Leighton et al.). Reviewing the information in Table 2, examinee three has examinee attributes of A1, A2, and A3 as noted by the binary values (111000), meaning the theoretical examinee that has mastered those three attributes. Given that information, it is expected that an examinee with those three attributes would correctly answer items 1, 2 and 3 as seen for that row in the expected response matrix (111000000000000), consequently, the total score would be three. Another example would be examinee 12 has A1, A2, A3, A4 and A6 identified in the binary values (111101), therefore, the expected response pattern for this examinee (111111000111000) indicates having these attributes, the examinee would correctly answer items 1, 2, 3, 4, 5, 6, 10, 11, and 12 correctly, resulting in a total score of nine. Being derived from the hierarchy, and therefore free of examinee error, AHM utilizes the expected examinees to estimate item parameters using item response theory (IRT) models (Leighton et al.).

Table 2. Expected Response Matrix, Total Scores and Examinee Attributes for a Hypothetical Set of 15 Examinees from Leighton, Geirl and Hunka, 2004.

Examinee	Expected Attributes	Expected Response Matrix	Total Score
1	100000	100000000000000	1
2	110000	110000000000000	2
3	111000	111000000000000	3
4	100100	100100000000000	2
5	110100	110110000000000	4
6	111100	111111000000000	6
7	100110	100100100000000	3
8	110110	110110110000000	6
9	111110	111111111000000	9
10	100101	100100000100000	3
11	110101	110110000110000	6
12	111101	111111000111000	9
13	100111	100100100100100	5
14	110111	110110110110110	10
15	111111	111111111111111	15

The expected item characteristic curves for each item are created using IRT under the assumption that the examinees' response correspond with the attribute hierarchy (Leighton et al.). The expected item characteristic curves are calculated using the item parameters dependent on the IRT model; using the ability parameter (Θ), item discrimination parameter (a_i), item difficulty parameter (b_i) for a two-parameter (2PL) logistic IRT model, adding a guessing parameter (c_i) for the three-parameter (3PL) logistic IRT model (Leighton et al.). The item parameters are calculated for each item based on the expected response patterns and the expected item characteristic curves, expected item and test information functions can be plotted (Leighton et al). Using person-fit indices, a researcher applying the AHM can then evaluate the extent to which the observed examinee response patterns align with the probability of correct response (Leighton et al.).

AHM requires the identification of an attribute hierarchy that is then seen throughout the matrices of the model, used to guide item development, and allows for the expected examinee response patterns to be defined and compared to observed examinee responses (Leighton et al., 2004). This process improves upon traditional interpretations of fit indices because the established link between a response pattern and the identified hierarchy; thereby supporting the identification of incongruent observed responses patterns, those patterns that deviate from the expected examinee response patterns (Leighton et al.). The differences are identified as the resulting pattern from the process results in -1, 0, and +1 values, where dj = 0 (no error), dj = -1 (error of 0 → 1, probability = $P_{jk}(\Theta)$, the probability of a correct response when an incorrect response was expected), and dj = +1 (error of 1 → 0, probability = 1- $P_{jm}(\Theta)$, the probability of an incorrect response when a correct response was expected) (Leighton et al.). To calculate the estimate for the likelihood that an observed response pattern approximates the expected examinee response pattern at a given ability parameter

$$P_{jExpected}(\Theta) = \prod_{k=1}^{K} P_{jk}(\Theta) \prod_{m=1}^{M} \left[1 - P_{jm}(\Theta)\right],$$

where k represents the subset of items with positive error (0 → 1) and m represent the subset of item with negative error (1 → 0) (Leighton et al.). Examinees are then classified as having a particular attribute by identifying the largest $P_{jExpected}(\Theta)$ (Leighton et al.).

Classification of examinees with AHM identifies the likely attribute combinations available for an examinee using one of two methods. The first method identifies the differences in the observed to expected response pattern and then uses the product of the probabilities of each difference to create the likelihood that the observed response pattern came from an expected response pattern for a specific ability level (Leighton et al.). The second method of classification involves identifying potential expected examinee response pattern matches within the observed response pattern and the associated attribute pattern is noted as present for the examinee (Leighton et al.). For the expected patterns not logically included, the likelihood of slips is concluded by identifying the slips and calculating the product of their probabilities, meaning the expected response pattern is aligned with the observed response pattern and only the ones in the expected response pattern are considered, making the errors in the comparison 1 → 0, which recall is the probability of an incorrect response when a correct response was expected (Leighton et al.).

The extension of AHM from the rule-space model was to develop a model that is capable of identifying the attributes each examinee has, and conversely, which attributes the examinee does not have, based on an attribute ierarchy (Leighton et al.) The value of this within education would be the potential to provide educators with precise feedback about student understanding, including identifying content a student has mastered and importantly, content the student is struggling with. Because AHM assumes attribute dependence, whereas rule-space model does not require attribute dependence for estimation, the essential assumption of AHM is the presence of an attribute hierarchy; therefore, test performance and classification accuracy is dependent on the correct identification of the hierarchy. It is this assumption that outlines the value and necessity of statistically identifying the presence of an attribute hierarchy. Consequently, the focus of this study is in evaluating the accuracy of a general diagnostic model, the Log-linear Cognitive Diagnostic Model (LCDM), with potential hierarchical data. However, because latent class based models assume all attribute profiles are present, there is concern for applying the models to data where a hierarchy is suspected (Templin & Bradshaw, 2013). Consequently, this study also intends to evaluate the efficiency of an adapted model, the Hierarchical Diagnostic

Classification Model (HDCM) as a means of addressing overfitting that can occur in general CDMs when an attribute hierarchy is present.

Log-linear Cognitive Diagnostic Model

Log-linear models with latent variables are models that define the probability of a correct response through the log-odds of a correct response for each item (Henson et al., 2009). von Davier (2005a) presents a general class of models for cognitive diagnosis (GDM) that is based on the extension, and designed to maintain similarities, of several previous models including, latent class models, item response theory models, the Rasch model and skill profile models. GDMs are extremely flexible within skill profile models as the model is capable of specifying both compensatory and noncompensatory (von Davier, 2005a). von Davier (2005a) defines the model using the skill profile, more precisely defined as a multidimensional latent variable, θ = ($a_{1}, \ldots, a_{K}$) and the user-defined skill levels $a_k \epsilon \{s_{k1}, \ldots, s_{kl}, \ldots, s_{kLk}\}$.

The Log-linear Cognitive Diagnostic Model (LCDM) is an extension of the GDM that allows for both compensatory and noncompensatory structures and uses dichotomous latent variables with dichotomous responses (Henson et al., 2009; Templin & Bradshaw, 2013). Much like the previously described models, the LCDM can be used to describe the conditional relationship between attributes and item response probability; however, LCDM does not mandate specifying models such as conjunctive or disjunctive (Henson et al.). Similar to factorial ANOVA, all attributes are crossed factors in the LCDM and are assumed as possible (Templin & Bradshaw). The fully crossed LCDM creates flexibility when evaluating the potential number of attributes that could potentially be measured by an item. The LCDM can be specified to model any number of attributes per item, however, consideration must be given to computation resources and time which may in turn limit the number of attributes being measured by each item (Templin & Bradshaw). The general form of the LCDM is as follows

$$P(X_{ei} = 1 \mid \alpha_e = \alpha_c) = \frac{\exp(\lambda_{i,0} + \lambda_i^T h(\alpha_e, q_i))}{1 + \exp(\lambda_{i,0} + \lambda_i^T h(\alpha_e, q_i))},$$

where q represents the Q-matrix entries for item i, the intercept parameter ($\lambda_{i,0}$) represents the log-odds of a correct response for an examinee who is not a master of the attribute(s) being measured, the attribute pattern (α_e) identifies examinees that have mastered ($\alpha_e = 1$) and those that have not mastered ($\alpha_e = 0$) the attribute being measured by the items, and the main effect (e.g., $\lambda_{i,1,(\mathrm{a})}$) represents the increase in log-odds given mastery of an attribute (Templin & Bradshaw). Because the number of possible attribute profiles in the LCDM is 2^A , vector $\boldsymbol{\lambda}_i$ represents a $(2^A - 1) \times 1$ vector of weights which is the LCDM parameters for item i while the function $\boldsymbol{h(q_i,\alpha_e)}$ is the vector-valued function of $(2^A - 1) \times 1$ that represents whether or not a parameter is present in an item the response function is expressed as:

$$\lambda_i^T h(\alpha_e, q_i) = \sum_{a=1}^{A} \lambda_{i,1,(a)} \alpha_{ea}, q_{ia} + \sum_{a=1}^{A-1} \sum_{b>a} \lambda_{i,2,(a,b)} \alpha_{ea} \alpha_{eb} q_{ia} q_{ib} + \ldots..$$

For an item i measuring two or more attributes, the response function includes all main effects ($\lambda_{i,1,(\mathrm{a})}$) and interactions between the attributes (e.g., $\lambda_{i,2,(\mathrm{a,b})}$, $\lambda_{i,1,(\mathrm{a,c})}$) or $\lambda_{i,1,(\mathrm{b,c})}$ are representative of two-way interactions between attribute a, b and c). The Q-matrix is then representative of these parameters; with the first elements of $\boldsymbol{h(q_i,\alpha_e)}$ being indicators of the main effect parameters and the second set of elements are indicators of all possible two-way interactions (Templin & Bradshaw).The interaction elements are products of multiplying the two attributes measured by the item with entries in the Q-matrix; for example, attributes 1 and 2 are measured by an item, the interaction indicator of these two attributes is represented by the product of α_{e1} × α_{e2} × q_{i1} × q_{i2} (Templin & Bradshaw). Should there be items measuring three attributes or more, the remaining linear combinations of $\boldsymbol{\lambda_i^T h(q_i, \alpha_e)}$ represent all remaining interactions (Templin & Bradshaw).

LCDM is a constrained model of a more general model that contains the attribute distribution within the base-rate parameters (π_c) to define the probability that a given examinee from a population has a given attribute pattern c (c = 1, . . . , 2^A) (Henson et al., 2009; Templin & Bradshaw, 2013). These combined with a measurement model to produce the marginal LCDM likelihood function for binary items and attributes for an examinee as follows:

$$P(X_e) = \sum_{c=1}^{2A} \pi_c \prod_{i=1}^{I} P(X_{ei} = 1 \mid \alpha_e)^{X_{ei}} (1 - P(X_{ei} = 1 | \alpha_e))^{1-X_{ei}}$$

where π_c is the probability an examinee will have attribute c and $2^A\pi_c$ represent the parameters of the joint distribution of the attributes. Templin and Bradshaw (2013) noted that the Multivariate Bernoulli Distribution (MVD) creates a categorical distribution that maps onto two-category latent attributes that has 2^A – 1 estimated π_c parameters under the LCDM.

Hierarchical Diagnostic Classification Model

Because the fully crossed LCDM assumes all combinations of attributes are present in the population the model estimates all possible patterns of mastery; however, when there is an attribute hierarchy the model over-fits the data by creating redundant item parameters and allowing classes to exist that are not present (Templin & Bradshaw, 2013). Templin and Bradshaw developed the Hierarchical Diagnostic Classification Model (HDCM) in response to the lack of statistical hypothesis tests for the presence of an attribute hierarchy. The HDCM is utilized as a constrained model of the fully crossed LCDM where an attribute hierarchy structure is present and is used to fix the redundant parameters created by the LCDM to zero (Templin & Bradshaw). Under LCDM, the number of attribute profiles estimated would be 2^A, while HDCM estimates only A + 1 attribute profiles (Templin & Bradshaw). For example, an educational assessment that measures four attributes would have $2^4 = 16$ possible attribute profiles under the fully crossed LCDM; however, under the HDCM there would only be 5 profiles estimated.

Under the HDCM model, when an attribute hierarchy is present, attributes represent *nested* factors wherein attributes that are dependent on another attribute are *nested* within the dependent attribute (Templin & Bradshaw, 2013). As noted under the LCDM, the attributes in profile *c* are represented as α_c where $c = 1, \ldots, 2^A$, but under HDCM the attributes in profile *c* are represented as $\alpha^*{}_c$ with the profiles identified as not possible removed (Templin & Bradshaw). HDCM is a constrained version of the LCDM, consequently, when there is a linear hierarchy, the set of item parameters need to reflect the nested structure of the attribute profiles, so the matrix portion of item response function

$$P(X_{ei} = 1 \mid \alpha_e = \alpha_c) = \frac{\exp(\lambda_{i,0} + \lambda_i^T h(\alpha_e, q_i))}{1 + \exp(\lambda_{i,0} + \lambda_i^T h(\alpha_e, q_i))},$$

becomes

$$\lambda_i^T h(\alpha_e^*, q_i) = \lambda_{i,1,(a)} \alpha_{ea} q_{ia} + \lambda_{i,2,(b(a))} \alpha_{ea} \alpha_{eb} q_{ia} q_{ib} \\ + \lambda_{i,3,(c(b,a))} \alpha_{ea} \alpha_{eb} \alpha_{ec} q_{ia} q_{ib} q_{ic} + \ldots$$

Templin and Bradshaw (2013) provide a simple example of an item that measures two attributes (*a* and *b*) with attribute *b* nested within attribute *a*, the shift in the item response function for examinee *e* on item *i* is as follows

$$P(X_{ei}=1 \mid \alpha_e=\alpha_c)=\frac{\exp(\lambda_{i,0}+\lambda_{i,1,(a)}\alpha_{ea}+\lambda_{i,2,(b(a))}\alpha_{ea}\alpha_{eb})}{1+\exp(\lambda_{i,0}+\lambda_{i,1,(a)}\alpha_{ea}+\lambda_{i,2,(b(a))}\alpha_{ea}\alpha_{eb})}.$$

For an item that measures two attributes, one being nested, the response function under the HDCM provides three parameters including the intercept, the main effect for the non-nested attribute and the interaction for the attribute nested within the first attribute (Templin & Bradshaw).

In addition to the change in the item parameters, the structural model is reduced as well, given that the parameters that represented attribute profiles that were not possible have been removed. There is a change in the base-rate parameters of 2^A for the LCDM to the reduced set of base-rate parameters that in turn creates a reduction in complexity of the structural by reducing the number of parameters under the HDCM (Templin & Bradshaw, 2013).

This study will evaluate the HDCM in an empirical setting involving a formative mathematics assessment designed using AHM; and is intended to evaluate the efficiency of HDCM as well as highlight the need for means to statistically evaluate the presence of an attribute hierarchy, therein extending Templin and Bradshaw’s seminal work into common K12 educational assessment.

CHAPTER THREE

Research Design

Overview

Within education, Cognitive Diagnosis Models (CDMs) provide information to support feedback regarding student strengths and weaknesses in content material. CDMs are capable of identifying specific skills a student has not mastered by linking item response performance with attributes (or skills) measured by the items on an assessment. For example, a diagnostic assessment built to identify areas a student is struggling with in mathematics not only provides an educator with the number of items the student answered correctly, but what skills the student has mastered or not mastered that were linked to the probability of a correct response on the missed items. The information resulting from CDMs can then be used to provide evidence of hierarchies, tailor student instruction and plan remediation. Though these models provide extensive application possibilities, should a hierarchy be present, conventional CDMs over fit the data by estimating all possible patterns of attribute mastery (2^k), therein specifying item parameters that are redundant (Templin & Bradshaw, 2013). For example, Table 3 is an example of an assessment measuring four attributes (ABCD) resulting in the estimation of 16 attribute profiles where a one represents the skill is included in the profile and a zero means the attribute is not included in the profile. Should two of those attributes be dependent on the mastery of other attributes being measured, a hierarchy exists, and there are now item parameters estimated for profiles that are not possible. Using the assessment example in Table 3, should attribute C be dependent on attribute A, the profiles that include attribute C without also including attribute A would not be possible based on that dependency.

Table 3. Example assessment measuring four attributes, resulting in 16 attribute profiles under the LCDM.

	Skills			
Profile	A	B	C	D
1	0	0	0	0
2	1	0	0	0
3	0	1	0	0
4	0	0	1	0
5	0	0	0	1
6	1	1	0	0
7	1	0	1	0
8	1	0	0	1
9	0	1	1	0
10	0	1	0	1
11	0	0	1	1
12	1	1	1	0
13	1	1	0	1
14	1	0	1	1
15	0	1	1	1
16	1	1	1	1

To address this over-specification, Templin and Bradshaw (2013) extended the Log-linear Cognitive Diagnostic Model (LCDM) to address cases where attribute hierarchies are present, therein establishing a link between LCDM and the functionality of the Attribute Hierarchy Model. Their work outlines the Hierarchical Diagnostic Classification Model (HDCM) as a probable solution to detecting the presence of attribute hierarchies that is absent in other latent class diagnostic models. Revisiting the previously mentioned four attribute assessment that resulted in 16 attribute profiles (2^4) using LCDM, assuming there are two attributes that are dependent on mastery of other attributes (e.g., attribute D is dependent on attribute C which is dependent on attribute A) an attribute hierarchy is now assumed to be present. Consequently, under LCDM, item parameters for attribute C and D have been estimated although, based on the hierarchy the attributes are dependent on others and the profiles for attribute C and D are not possible. HDCM can be used to form a hypothesis test to evaluate the presence of a potential attribute hierarchy, influencing the understanding of the attributes that are measured by each test, creating an empirical evaluation of potential hierarchies (Templin & Bradshaw). In addition to simulation work to establish the model's detection ability, Templin and Bradshaw (2013) evaluated HDCM with an empirical example for English Language Proficiency statistically establishing the presence of an attribute hierarchy. This study is designed to extend that work to evaluate the estimation ability of the model in common K-12 standards-based assessments by using a formative mathematics assessment data set based on a theorized learning hierarchy. Using HDCM after LCDM allows a researcher to investigate and identify the presence of the hierarchy in tandem with parameter estimations. Because there is a proposed hierarchy, using HDCM as a follow-up to the LCDM allows the researcher to create the nested structure and rerun the analysis and statistically determine if the hierarchy is present in the data. By evaluating the ability of a model to statistically identify a learning hierarchy, this study could potentially provide supportive information to an educator about the progression of learning and when skills should be taught or revisited.

Hypothesis

There is an attribute hierarchy present in how students learn and understand the essential concepts of slope; consequently, the HDCM will fit the data from a formative mathematics assessment designed to measure the hierarchy better than the LCDM.

Research Questions

1. Using a formative assessment designed to measure a specific and well-defined learning hierarchy, does the HDCM accurately identify the presence or absence of the hierarchy?
2. What are the estimations and classification differences between the HDCM, LCDM and AHM from a formative assessment designed using AHM to measure a learning hierarchy?

Assessment

The Foundational Concepts of Slope Assessment (Broaddus, 2011) was a formative assessment, designed to measure the understanding of slope, based on the attributes of the Foundational Concepts of Slope Attribute Hierarchy (FCSAH). The theorized learning hierarchy is based on extensive previous theory, research, and subject matter expertise regarding how students learn and understand covariation and proportional reasoning necessary for continued learning and understanding of slope (Broaddus, 2011). During her work, Broaddus aligned five attributes in the learning order at which attributes might optimally be acquired. A summary of the attributes is presented in Table 4. The theorized learning progression of the FCASH indicates that a student must first master attribute one, followed by attribute two, and then may master attribute three, four, or five in any order (Broaddus). This hierarchy can be seen in Figure 2.

Table 4. Summary of the Attributes of the FCSAH

Attribute A1	Detect which quantities in a problem situation varied in correspondence to one another without any reference to their directions of change
Attribute A2	Identify the direction of change of two covariates in constant rate problem contexts
Attribute A3	Interpret the meaning of slope ratio in terms of the context of a problem presented either verbally or graphically concerning slopes whose ratio values simplified to whole numbers
Attribute A4	Interpret the meaning of slope ratio in terms of the context of a problem presented either verbally or graphically concerning slopes whose ratio values simplified to unit fractions
Attribute A5	Interpret the meaning of slope ratio in terms of the context of a problem presented either verbally or graphically concerning slopes whose ratio values simplified to positive rational numbers but neither whole numbers nor unit fractions

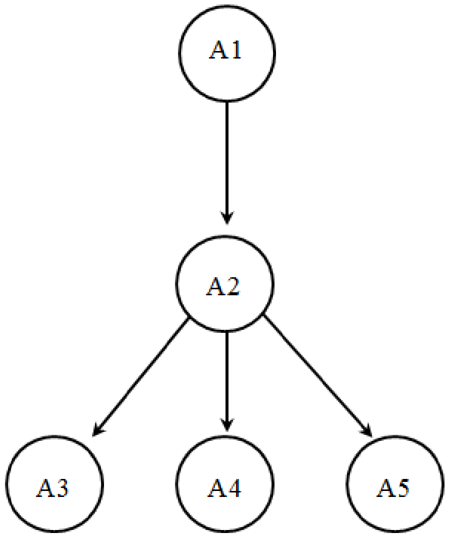


Figure 2. The diagram graphically represents the structure of the FCSAH. As shown, a student must master attribute one (A1) before mastering attribute two (A2), from which the student can master any combination of attributes three (A3), attribute four (A4) or attribute five (A5), and this entire pattern must be followed in sequence.

Broaddus analyzed the FCSAH and developed matrix representations of the attributes with their dependent hierarchical combinations. Broaddus (2011) used these matrices to develop test items specifically targeted to the attributes contained in the FCSAH. The adjacency matrix was created to evaluate the relationships of attributes in the attribute hierarchy using ones and zeros (Matrix 5). The reachability matrix was used to depict direct and indirect relationships among the attributes in the attribute hierarchy (Matrix 6), while the incidence matric represented all possible combinations of the attributes in a hierarchy (Matrix 7). Finally, Broaddus created a reduced incidence matrix to represent the possible combinations of attributes that meet the constraints defined by the attribute hierarchy (Matrix 8).

Matrix 5: Adjacency Matrix for the FCSAH

$$\begin{array}{cccccc}
 & A1 & A2 & A3 & A4 & A5 \\
A1 & 0 & 1 & 0 & 0 & 0 \\
A2 & 0 & 0 & 1 & 1 & 1 \\
A3 & 0 & 0 & 0 & 0 & 0 \\
A4 & 0 & 0 & 0 & 0 & 0 \\
A5 & 0 & 0 & 0 & 0 & 0
\end{array}$$

Matrix 6: Reachability Matrix for the FCSAH

$$\begin{array}{cccccc}
 & A1 & A2 & A3 & A4 & A5 \\
A1 & 1 & 1 & 1 & 1 & 1 \\
A2 & 0 & 1 & 1 & 1 & 1 \\
A3 & 0 & 0 & 1 & 0 & 0 \\
A4 & 0 & 0 & 0 & 1 & 0 \\
A5 & 0 & 0 & 0 & 0 & 1
\end{array}$$

Matrix 7: Incidence Matrix for the FCSAH

$$\begin{array}{ccccccccccccccccccccccccccccccccc}
A1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\
A2 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\
A3 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\
A4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
A5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1
\end{array}$$

Matrix 8: Reduced Incidence Matrix for the FCSAH

$$
\begin{array}{cccccc}
 & T1 & T2 & T3 & T4 & T5 \\
A1 & 1 & 1 & 1 & 1 & 1 \\
A2 & 0 & 1 & 1 & 1 & 1 \\
A3 & 0 & 0 & 1 & 0 & 0 \\
A4 & 0 & 0 & 0 & 1 & 0 \\
A5 & 0 & 0 & 0 & 0 & 1
\end{array}
$$

The FCSA was assembled and the expected item response vectors were created to represent the expected student responses dependent on which attributes from the FCSAH they had mastered (Broaddus, 2011). These vectors were configured with each row representing a potential answer pattern, while the matrix represented all potential attribute mastery combinations in the hierarchy. The item and test design, as well as eventual score interpretation, was informed using a reduced incidence matrix (Qr). This matrix contained five columns representative of the linearity of the FCSAH and five lines, one for each of the attributes of the FCASH (Broaddus).

The FCSA was designed to include four items per attribute, resulting in a 20-item assessment. For attribute 1 (A1), the four test items were all word items and the researcher determined ordering based on item difficulty was optimal for this attribute item set (Broaddus, 2011). The four items designed to measure Attribute 2 (A2) contained graphs or verbal descriptions, alternating item sequence so that within the assessment the student would complete an item with a graph, then an item with a verbal description. This sequencing pattern was continued for Attributes 3, 4 and 5 as well. By utilizing item design choices of word problems, graph inclusion and verbal descriptions, the researcher was able to evaluate the student’s ability to perceive covariation across problem contexts (Broaddus). The remaining three attributes were concerned with a student's proportional reasoning abilities using whole numbers, unit fractions, and positive rational numbers that are not whole numbers nor unit fractions. For attribute 3 (A3), the four items presented problem contexts verbally and graphical for slopes with simplified ratio values of whole numbers (Broaddus). The four items of attribute 4 (A4) presented problem contexts verbally and graphical for slopes whose ratio values simplified to unit fractions (Broaddus). Attribute 5 (A5) was measured using four items that presented problem contexts verbally and graphical for slopes whose ratio values simplified to positive rational numbers, however, to separate from A3 and A4, the simplified ratio values were neither whole nor unit fractions (Broaddus).

Broaddus used the student responses to estimate the item parameters of the twenty items of the FCSA, the item parameters from her study are shown in Table 5. These parameters were then used to produce the item characteristic curve (ICC), highlighting the relationship between item responses and the ability needed to correctly answer an item (Broaddus). Her analysis found that the items of the FCSA discriminated well between students of higher and lower ability levels, though nineteen of the twenty items had relatively low difficulty. Summing the ICCs, the test characteristic indicated that the FCSA was not very difficult as the point of inflection has an abscissa between -1 and 0, nor did the FCSA differentiate students of different ability levels as the slope of the test characteristic was not steep (Broaddus). The item information function, representing the amount of information the item contributes for the different ability levels of the students in the data set, highlighted that though many of the items on the FCSA were informative for students within an ability range of -1.6 to 0.0, many of the items were answered correctly by a large number of students and were not informative (Broaddus). Finally, the test information function, representing the summation of the item information functions, noted that the FCSA was most informative for students within the ability range of -2.0 and 1.0.

Table 5. Item Parameter Estimates for the FCSA

Item	a-parameter	b-parameter	c-parameter
1	0.48	-2.35	0.26
2	0.69	-2.95	0.24
3	0.58	-2.03	0.24
4	0.65	-1.47	0.26
5	0.73	-2.06	0.26
6	0.74	-1.17	0.31
7	0.87	0.00	0.27
8	0.67	-1.15	0.28
9	0.71	-0.76	0.25
10	1.41	-0.82	0.23
11	1.00	-0.43	0.19
12	1.24	-1.27	0.19
13	1.01	-0.03	0.14
14	1.10	-0.95	0.15
15	0.89	-0.08	0.18
16	0.65	-1.51	0.23
17	0.76	-1.23	0.17
18	1.36	-1.20	0.22
19	1.03	-0.24	0.20
20	0.82	2.09	0.26

Data

The FCSA was delivered online to 1629 students in middle and high schools who were enrolled in Pre-Algebra, Algebra1, Geometry, Algebra 2 or similar courses that are typically taken before pre-Calculus. Student responses were captured by a state assessment delivery engine and were configured within an excel datasheet that contained the course name, student's district number, the student's responses to the 20 items on the assessment recoded to binary data, with 0 representing an incorrect response and 1 being correct response, and the student's overall percent correct (Broaddus, 2011).

Procedures

This study evaluated the estimation and classification accuracy of the LCDM and the HDCM within a mathematics assessment data set. The initial analysis used the LCDM to determine classification and model fit. Subsequent analysis used the HDCM to determine the presence of potential attribute hierarchy by creating nested attributes based on the initial output. As part of a parameter recovery study, item parameters, the structural model parameters and the classification accuracy of examinee attribute profiles for the two models were evaluated.

Q-Matrix

The Q-matrix used in this study was the result of psychometric mapping done by Broaddus (2011), showing that the twenty items of the FCSA were likely measuring five skills. Table 6 shows the Q-matrix mapping each of the five skills in the 20-item FCSA. For example, items 13, 14, 15, and 16 measure attribute 4 (A4), which would be items measuring the student’s ability to interpret the meaning of slope ratio in terms of the context of a problem presented either verbally or graphically concerning slopes whose ratio values simplified to unit fractions.

Table 6. FCSA Q-Matrix

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

Software

MPlus7 was used to estimate both LCDM and HDCM, while R was used to calculate significance of the deviance statistic.

LCDM Analysis

Estimation

The structural parameter estimates were used to outline the probability a student would have a given attribute profile, indicating the probability of membership in a latent class (Rupp et al., 2010). As there are five attributes being measured in the FCSA, there are 32 latent classes ($2^A = 2^5 = 32$), representing attribute mastery. The structural parameter estimates were used to determine classification and evaluate model fit within the LCDM analysis, allowing me to in turn evaluate the data for a potential hierarchy under the HCDM. MPlus calculates the latent class membership probabilities estimates (υ_c) by transforming the μ_c parameters:

$$\upsilon_c = \frac{\exp(\mu_c)}{\sum_{c=1}^{C}\exp(\mu_c)} \tag{1}$$

Item parameters provided the probability that an examinee in a given latent class would answer an item correctly. Item parameters are valuable for identifying student learning and teaching paths as identifying a hierarchy could highlight the need to teach a given skill prior to the remaining (Templin & Bradshaw, 2013). Rupp et al. (2010) noted the equation for the general latent class model is

$$P(X_r = x_r) = \sum_{c=1}^{C} \upsilon_c \prod_{i=1}^{I} \pi_{ic}^{x_{ir}} (1-\pi_{ic})^{1-x_{ir}} \tag{2}$$

where the probability (*P*) and observed response data across items (X_r and x_r) is a function of the probability of latent class membership (υ_c), probability of a correct response to item *i* dependent on latent class membership (π_{ic}) and the observed response (x_r).

LCDM Model Specification

Using the algorithm defined by Rupp et al. (2010), a class-to-profile table was created to define the unique attribute profile for each latent class or mastery profile. Because there were five attributes being measured with the FCSA, under the LCDM there would be $2^5 = 32$ possible attribute profiles to indicate if a student had mastered or had not mastered an attribute, representing all possible attribute mastery/non-mastery combinations for the five attributes (Table 7).

Table 7. Attribute Mastery Profiles Under the LCDM

Mastery profile	Attribute 1	Attribute 2	Attribute 3	Attribute 4	Attribute 5
α_{e1}	0	0	0	0	0
α_{e2}	0	0	0	0	1
α_{e3}	0	0	0	1	0
α_{e4}	0	0	0	1	1
α_{e5}	0	0	1	0	0
α_{e6}	0	0	1	0	1
α_{e7}	0	0	1	1	0
α_{e8}	0	0	1	1	1
α_{e9}	0	1	0	0	0
α_{e10}	0	1	0	0	1
α_{e11}	0	1	0	1	0
α_{e12}	0	1	0	1	1
α_{e13}	0	1	1	0	0
α_{e14}	0	1	1	0	1
α_{e15}	0	1	1	1	0
α_{e16}	0	1	1	1	1
α_{e17}	1	0	0	0	0
α_{e18}	1	0	0	0	1
α_{e19}	1	0	0	1	0
α_{e20}	1	0	0	1	1
α_{e21}	1	0	1	0	0
α_{e22}	1	0	1	0	1
α_{e23}	1	0	1	1	0
α_{e24}	1	0	1	1	1
α_{e25}	1	1	0	0	0
α_{e26}	1	1	0	0	1
α_{e27}	1	1	0	1	0
α_{e28}	1	1	0	1	1
α_{e29}	1	1	1	0	0
α_{e30}	1	1	1	0	1
α_{e31}	1	1	1	1	0
α_{e32}	1	1	1	1	1

Note: mastery = 1; non-mastery = 0

The class-to-profile table defined the LCDM parameterization for each item response probability (π_{ic}). The LCDM parameterization and the Q-matrix were used to create the item response functions for each item parameter (π_{ic}). This resulted in two possible item response functions per item: the intercept parameter for items that do not measure an attribute and the main effect parameter for items that do measure a given attribute. Having created the class-to-profile table and defining the item response parameters, the LCDM parameters were identified for items and classes, therein identifying the LCDM kernels for each item and latent class. For each item, the LCDM and the latent class defined the threshold value used in the program (Rupp et al., 2010).

Model Fit

The global fit statistics for the model was used to assess fit of the LCDM as a non-nested model including the log-likelihood, information criteria and the chi-square test of model fit. The chi-square test of model fit outlines global fit, providing Pearson Chi-Square and the Likelihood ratio Chi-Square test. Because the FCSA has 20 items, the expected possible response patterns is quite large, consequently, goodness-of-fit was evaluated using pairs of items, looking at the observed and model-predicted values, as well as the Bivariate Pearson Chi-Square and Bivariate Log-Likelihood Chi-Square (Rupp et al., 2010).

Parameters Evaluation

The estimated number of examinees in each latent class was evaluated, including the proportion conversion (υ_c), the estimated values for μ_c was used to compute the values of υ_c, and the estimates for the LCDM parameters and log-odds were used to calculate the probability of getting an item correct (π_{ic}). Finally, the estimated student parameters were investigated. This section of the MPlus output provided information about the examinee’s responses, the posterior probabilities by attribute profile, and the maximum a posteriori estimates.

Hierarchy Evaluation

The examinee classifications, as well as the item parameter estimates from the full LCDM, were reviewed for evidence of a hierarchy by looking at the proportion of examinees for each attribute profile. It is assumed that small proportions would highlight potential learning hierarchies (Gierl, Leighton, & Hunka, 2007; Templin & Bradshaw, 2013).

HDCM Analysis

The HDCM analysis followed along similar lines as the full LCDM analysis wherein the number of latent attributes and the loading patterns that were determined in the LCDM analysis were used. The principal model difference of HDCM from LCDM is the model does not enforce all attribute combination estimations in the presence of an attribute hierarchy, as the structural model is reduced because some attribute profiles cannot occur. Consequently, the full LCDM analysis was utilized to investigate the presence of a potential attribute hierarchy and model fit, modifying the parameter settings to align to the FCSAH, using the HDCM, shifting from 2^A base-rate parameters to A + 1.

Model Comparison

The item and structural model parameter estimates were evaluated for the full LCDM and the nested-attribute HDCM, noting differences in the model fit, intercept and main effect values and evaluated the examinee classification estimates to assess model agreement. Theoretically, based on Templin and Bradshaw’s previous work, when a hierarchy is present, the LCDM attempts to use too many parameters to fit the values. The HDCM creates nested attributes, eliminating redundant profiles; consequently, the item parameter estimates should be more stable, having smaller error values and improving model fit. After investigating model comparability, the hypothesis test was used to determine if an attribute hierarchy was present using the naïve distribution, understanding that the distribution is overly conservative.

In addition to item parameter estimates, the examinee classifications for agreement between the full LCDM and the HDCM were evaluated. Templin and Bradshaw (2013) found high agreement between the two models in previous research using language assessment results; It is important to determine if that pattern of agreement exists with common K12 formative data. Additionally, by comparing these findings to a previous Attribute Hierarchy Model analysis, differences in the attribute mastery probabilities determined by the models were identified.

Significance of Study

There is a common understanding that the progression of learning occurs in steps, requiring a student to learn one concept prior to being able to complete another; consequently, there is a sequential order to most education processes. For example, in mathematics, a student must know how to subtract before being able to successfully complete division problems. While understanding the concept of nouns and verbs is required before a student can complete sentence construction tasks. The theories underlying curriculum creation support sequential learning, building attribute hierarchies throughout learning, therefore, student responses should represent current understanding and provide insight into the resulting hierarchy (Templin & Bradshaw, 2013). As noted in Templin and Bradshaw's (2013) work, should researchers utilize a nonhierarchical DCM when an attribute hierarchy is present, the model over-fits the data specifying item parameters and allowing for mastery profiles that are not present due to the attribute hierarchies. This study utilized the HDCM to investigate the underlying structure of the attributes measured by a formative assessment, potentially identifying content hierarchies. Additionally, as HDCM is a relatively new model, this study extends the research of HDCM by using mathematics data from a formative scenario and highlights the benefits of statistically identifying the presence of an attribute hierarchy prior to subsequent student classification analysis.

CHAPTER FOUR

Results

Introduction

Within education, Cognitive Diagnosis Models (CDMs) provide explicit feedback about student understanding, including areas of strength and areas where a student might be struggling with content. Because CDMs are capable of identifying specific skills a student has not mastered by linking item response performance with attributes (or skills) measured by the items on an assessment, the models provide information to support evaluation and revision of education plans. By identifying areas a student is struggling in, the information resulting from CDMs can be used to provide evidence of hierarchies, tailor student instruction and create remediation plans. As previously noted, these models provide extensive application possibilities, however, when there is a learning hierarchy present, conventional CDMs over fit the data by estimating all possible patterns of attribute mastery (2^k), meaning item parameters that are redundant and do not exist are still specified (Templin & Bradshaw, 2013). For example, an assessment measuring five attributes would result in the estimation of 32 attribute profiles; however, should a hierarchy exists where two of the five attributes are dependent on the mastery of another attribute being measured, there are now item parameters estimated for profiles that are not possible.

Templin and Bradshaw (2013) extended the Log-linear Cognitive Diagnostic Model (LCDM) to address this over-specification when attribute hierarchies are present; establishing a link between LCDM and the functionality found in the Attribute Hierarchy Model (AHM). Their work outlines the Hierarchical Diagnostic Classification Model (HDCM) as a solution to detecting the presence of attribute hierarchies that is absent in other latent class diagnostic models. Again, considering the previously mentioned five attribute assessment where two of the five are dependent on another measured skill would result in 32 attribute profiles (2^5) using LCDM. In this example, under LCDM, item parameters are estimated for all of the 32 attribute profiles even though several are not possible due to dependency on another skill. It is in this scenario where HDCM is most useful and can be used to evaluate the presence of a potential attribute hierarchy, influencing the understanding of the sequence of mastery for the attributes that are measured by a test, potentially identifying a hierarchy (Templin & Bradshaw). Templin and Bradshaw (2013) evaluated HDCM with an empirical example for English Language Proficiency, this study is designed to extend that work to evaluate the estimation ability of the model in a common K-12 standards-based assessment scenario by using a data set from a formative mathematics assessment designed to evaluate student understanding of slope. Using HDCM after LCDM allows a researcher to investigate the presence of a theorized attribute hierarchy in tandem with parameter estimations, in this case the FCSAH. Should a hierarchy be theorized initially or suspected from the LCDM analysis, using HDCM allows the researcher to create the nested structure and rerun the analysis to investigate further.

Hypothesis

There is an attribute hierarchy present in how students learn and understand the essential concepts of slope; consequently, the HDCM will fit the data from a formative mathematics assessment designed to measure the hierarchy better than the LCDM.

Research Questions

1. Using a formative assessment designed to measure a specific and well-defined learning hierarchy, does the HDCM accurately identify the presence or absence of the hierarchy?
2. What are the estimations and classification differences between the HDCM, LCDM and AHM from a formative assessment designed using AHM to measure a learning hierarchy?

This chapter presents the results of this study in three sections, beginning with a review of the assessment used in the study, the Foundational Concepts of Slope Assessment (FCSA), as well as a review of the design model, Attribute Hierarchy Model (AHM). The second section presents the results of the Log-linear Cognitive Diagnostic Model (LCDM), highlighting the structural and item classifications of the model output. The final section presents the results of the Hierarchical Diagnostic Classification Model (HDCM). This portion of the results identifies the constraints made to the LCDM syntax to align to the FCSAH which is theorized to be measured by the assessment and presents the statistical analysis used to evaluate the presence of an attribute hierarchy. This final section allows for a direct comparison of the data using HDCM against a previous AHM research study used to analyze the same data set.

FCSA Development and Slope Attribute Hierarchy

The FCSA served as a formative assessment, designed to measure the understanding of slope, based on the attributes of a theorized learning hierarchy, the Foundational Concepts of Slope Assessment Hierarchy (FCSAH). The theorized learning hierarchy is based on extensive previous theory, research, and subject matter expertise regarding how students learn and understand covariation and proportional reasoning necessary for continued learning and understanding of slope (Broaddus, 2011). From this work, Broaddus defined and aligned five attributes in an order of which attributes might optimally be acquired (Table 8). The theorized learning progressions of the FCASH indicates that a student must first master attribute one, followed by attribute two, and then may master attribute three, four, or five in any order (Broaddus). This hierarchy can be seen in Figure 2.

Table 8. Summary of the Attributes of the FCSAH

Attribute A1	Detect which quantities in a problem situation varied in correspondence to one another without any reference to their directions of change
Attribute A2	Identify the direction of change of two covariates in constant rate problem contexts
Attribute A3	Interpret the meaning of slope ratio in terms of the context of a problem presented either verbally or graphically concerning slopes whose ratio values simplified to whole numbers
Attribute A4	Interpret the meaning of slope ratio in terms of the context of a problem presented either verbally or graphically concerning slopes whose ratio values simplified to unit fractions
Attribute A5	Interpret the meaning of slope ratio in terms of the context of a problem presented either verbally or graphically concerning slopes whose ratio values simplified to positive rational numbers but neither whole numbers nor unit fractions

The FCSA was assembled and expected item response vectors were created to represent the expected student responses dependent on which attributes from the FCSAH they had mastered (Broaddus, 2011). These vectors were configured with each row representing a potential answer pattern, while the matrix represented all of the potential attribute mastery combinations in the hierarchy. The item and test design, as well as eventual score interpretation, was informed using a reduced incidence matrix (Qr). This matrix contained five columns representative of the linearity of the FCSAH and five lines, one for each of the attributes of the FCASH (Broaddus).

The FCSA was designed to include four items per attribute, resulting in a 20-item assessment. For attribute 1 (A1), the four test items were all word items and the researcher determined ordering based on item difficulty was optimal for this attribute item set (Broaddus, 2011). The four items designed to measure Attribute 2 (A2) contained graphs or verbal descriptions. By utilizing item design choices of word problems, graph inclusion and verbal descriptions, the researcher was able to evaluate the student's ability to perceive covariation across problem contexts (Broaddus). The remaining three attributes were concerned with a student’s proportional reasoning abilities using whole numbers, unit fractions, and positive rational numbers that are not whole numbers nor unit fractions. For attribute 3 (A3), the four items presented problem contexts verbally and graphical for slopes with simplified ratio values of whole numbers (Broaddus). The four items of attribute 4 (A4) presented problem contexts verbally and graphical for slopes whose ratio values simplified to unit fractions (Broaddus). Attribute 5 (A5) was measured using four items that presented problem contexts verbally and graphical for slopes whose ratio values simplified to positive rational numbers, however, to separate from A3 and A4, the simplified ratio values were neither whole nor unit fractions (Broaddus). For item order, the researcher chose to alternate the item sequence; presenting an item with verbal descriptions followed by an item that relied on graphical descriptions.

Data Set

The FCSA was delivered online to 1629 students in middle and high schools who were enrolled in Pre-Algebra, Algebra1, Geometry, Algebra 2 or similar courses that are typically taken before pre-Calculus. Student responses were captured by a state assessment delivery engine and were configured within an excel datasheet that contained the course name, student’s district number, the student’s responses to the 20 items on the assessment recoded to binary data, with 0 representing an incorrect response and 1 being correct response, and the student’s overall percent correct (Broaddus, 2011).

Previous AHM Analysis

The initial analysis of the FCSA data set was conducted using AHM. This analysis used the five attributes as the independent variables and the ten latent variables that represented the expected response vectors as the dependent variables (Broaddus, 2011). These ten expected student response vectors represent the differing combinations of attributes a hypothetical student could have possessed. These ten expected response vectors were then analyzed with item response theory (IRT) calculations to estimate the ability of each of the ten hypothetical students represented by the expected response vectors (Broaddus).

As noted, the observed student responses were converted to a binary file where ones represented correctly answered items and zeros for items the student answered incorrectly. Broaddus (2011) then compared each observed vector to every expected response vector to produce the likelihood the student had the same ability estimate as the expected response vector, resulting in ten likelihood estimates for each student in the data set. To determine the likelihood, Broaddus compared each observed student response vector to and subtracted from each expected response vector to produce a difference vector, resulting in values of 1, 0, or -1. A 1 represented a student error; an incorrect response from a student who should have answered the item correctly because the student had the knowledge implied by the expected response vector. An entry of 0 in the difference vector indicated that the student responded as expected; those with the knowledge implied by the expected response vector answered correctly, those without the knowledge implied by the expected response vector answered incorrectly. An entry of -1 indicated that the student with the knowledge implied by the response vector should have answered the item incorrectly but answered the item correctly. The difference vector was used in the formula (equation 3) to create the likelihood the student had the same ability estimate as the corresponding expected response vector, these ten likelihood estimates were summed for each student (Broaddus).

$$P_{j\,expected}(\theta) = \prod {}_{k=1}^{k} P_{jk}(\theta) \prod {}_{m=1}^{m} \left\lfloor 1 - P_{jm}(\theta) \right\rfloor \qquad (3)$$

For this comparison formula, Broaddus utilized the expected response vector compared to observed student response, the number of ones and negative ones in the difference vector to produce estimates of the likelihood that an observed student response vector matched the expected response vector. Each likelihood estimate was divided by the student sum to produce a probability corresponding to each likelihood estimate. The highest probability for each student was used to classify the student into the knowledge state represented by the expected response vector corresponding to the highest probability. Broaddus used these probabilities to evaluate the likelihood that the FCSAH was true.

The 1629 student participants were classified into one of the ten knowledge states dependent on the FCSAH, from which Broaddus utilized the ability estimates and the knowledge state classifications to evaluate further (Table 9). Based on her analysis, the three profiles that had the largest proportion of students were as follows: having mastered A1245 (21%), having masteredA1235 (19%), and having mastered A1234. Her evaluation found that students classified into knowledge state A0 appeared to have different ability levels than other students and students classified as A1 and A12 had similar ability levels. Those classified as A123, A124 and A125 shared very similar ability levels, but different than those in other knowledge states, which Broaddus interpreted as meaning the students being classified to the knowledge state A123, 124 and 125 did not offer much information. Those in A1234, 1235 and 1245 had similar ability levels, while those at 12345 had very different ability levels than students classified at lower levels.

Table 9. AHM Analysis Percent of Students by Mastery Profile

Profile	Profile Ability Estimate	Percent of Students
A0	-2.92	1
A1	-2.23	2
A12	-1.67	3
A123	-0.95	9
A124	-1.19	9
A125	-1.23	8
A1234	-0.14	16
A1235	-0.21	19
A1245	-0.42	21
A12345	1.45	13

LCDM results

Log-linear models with latent variables are models that define the probability of a correct response through the log-odds of a correct response for each item (Henson et al., 2009). The general class of models for cognitive diagnosis (GDM) presented by von Davier (2005a) is based on the extension, and designed to maintain similarities, of several previous models including, latent class models, item response theory models, the Rasch model and skill profile models. GDMs are extremely flexible within skill profile models as the model is capable of specifying both compensatory and noncompensatory (von Davier, 2005a). The Log-linear Cognitive Diagnostic Model (LCDM) is an extension of the GDM that maintains that flexibility and can be used to describe the conditional relationship between attributes and item response probability; however, LCDM does not mandate specifying models (Henson et al.). Comparable to factorial ANOVA, all attributes are crossed factors in the LCDM and are assumed as possible (Templin & Bradshaw). The fully crossed LCDM creates flexibility when evaluating the possible number of attributes that could potentially be measured by an item and can be specified to model any number of attributes per item (Templin & Bradshaw).

Mastery Profiles

For this study, there were five attributes being measured with the FCSA, consequently, under the LCDM there would be $2^5 = 32$ possible attribute profiles to indicate if a student had mastered or had not mastered an attribute, representing all possible attribute mastery/nonmastery combinations for the five attributes (Table 11).

Table 11. Attribute Mastery Profiles Under the LCDM

Mastery profile	Attribute 1	Attribute 2	Attribute 3	Attribute 4	Attribute 5
α_{e1}	0	0	0	0	0
α_{e2}	0	0	0	0	1
α_{e3}	0	0	0	1	0
α_{e4}	0	0	0	1	1
α_{e5}	0	0	1	0	0
α_{e6}	0	0	1	0	1
α_{e7}	0	0	1	1	0
α_{e8}	0	0	1	1	1
α_{e9}	0	1	0	0	0
α_{e10}	0	1	0	0	1
α_{e11}	0	1	0	1	0
α_{e12}	0	1	0	1	1
α_{e13}	0	1	1	0	0
α_{e14}	0	1	1	0	1
α_{e15}	0	1	1	1	0
α_{e16}	0	1	1	1	1
α_{e17}	1	0	0	0	0
α_{e18}	1	0	0	0	1
α_{e19}	1	0	0	1	0
α_{e20}	1	0	0	1	1
α_{e21}	1	0	1	0	0
α_{e22}	1	0	1	0	1
α_{e23}	1	0	1	1	0
α_{e24}	1	0	1	1	1
α_{e25}	1	1	0	0	0
α_{e26}	1	1	0	0	1
α_{e27}	1	1	0	1	0
α_{e28}	1	1	0	1	1
α_{e29}	1	1	1	0	0
α_{e30}	1	1	1	0	1
α_{e31}	1	1	1	1	0
α_{e32}	1	1	1	1	1

Note: mastery = 1; non-mastery = 0

The LCDM utilized the Q-matrix mapping of the five skills within the FCSAH to predict the item responses conditional on the student's mastery profile α_e = [α_{e1}, α_{e2}, $\alpha_{e3,}$...α_{e32}] using the item response function. Considering a specific example, item 12 in the FCSA is intended to measure attribute three, so the item response function would be as follows:

$$P(X_{ei} = 1 | \alpha_e) = \frac{exp(\lambda_{12,0} + \lambda_{12,1,(3)}\alpha_{e1})}{1 + exp(\lambda_{12,0} + \lambda_{12,1,(3)}\alpha_{e1})} \tag{4}$$

The Q-matrix used in this study was the result of psychometric mapping done by Broaddus (2011), showing that the twenty items of the FCSA were likely measuring five skills. The first six columns of Table 3 shows the Q-matrix mapping each of the five skills in the 20-item FCSA. For example, items 9, 10, 11, and 12 measure attribute 3 (A3), which would be items that present problem contexts verbally and graphical for slopes with simplified ratio values of whole numbers. It is this Q-matrix and the LCDM that was used to establish a fixed number of classes that is determined by the number of total possible attribute patterns and the fixed item parameter structure.

Model and Item Fit

The initial analysis focused on the item fit statistics; using the bivariate model fit information as the index of fit for each pair of items in the FCSA to determine if there are items that do not fit the model. By comparing the observed and the expected frequencies of responses, a chi-square value is provided to represent degree of misfit. This analysis highlighted item 18 as not well fit by the model, χ^2 = 16.15, p < .01, with an overall χ^2 = 363.73. Item 18 was removed, the model rerun and overall model fit increased as evidenced by a decrease in the overall bivariate chi-square value, χ^2 = 289.79. This 19-item model was retained for the remaining LCDM and the HDCM analyses.

Table 12. FCSA Q-Matrix and LCDM Item Parameter Estimates

	Skill	Item						
i	Measured	Difficulty	$\lambda_{i,0}$	$\lambda_{i,1}$	$\lambda_{i,2}$	$\lambda_{i,3}$	$\lambda_{i,4}$	$\lambda_{i,5}$
1	A1	-2.35	2.91(0.21)	2.01(0.26)				
2	A1	-2.95	5.04(0.67)	3.26(0.68)				
3	A1	-2.03	2.92(0.24)	2.12(0.28)				
4	A1	-1.47	2.90(0.28)	2.76(0.28)				
5	A2	-2.06	3.92(0.27)		2.74(0.31)			
6	A2	-1.17	2.65(0.18)		2.21(0.21)			
7	A2	0.00	1.33(0.11)		2.20(0.18)			
8	A2	-1.15	2.28(0.13)		1.86(0.18)			
9	A3	-0.76	1.90(0.11)			1.95(0.15)		
10	A3	-0.82	3.32(0.23)			3.56(0.25)		
11	A3	-0.43	1.70(0.11)			2.35(0.15)		
12	A3	-1.27	3.70(0.23)			3.24(0.26)		
13	A4	-0.03	1.11(0.10)				2.30(0.15)	
14	A4	-0.95	2.76(0.17)				2.75(0.19)	
15	A4	-0.08	1.18(0.10)				2.06(0.15)	
16	A4	-1.51	2.49(0.14)				1.72(0.18)	
17	A5	-1.23	2.60(0.18)					2.24(0.20)
18	A5	-1.20						
19	A5	-0.24	1.74(0.17)					2.58(0.22)
20	A5	2.09	-0.50(0.08)					0.58(0.15)

Note: Item 18 was removed in subsequent analysis.

Item Parameters

As there are five attributes being measured with the FCSA, under the fully crossed LCDM, there are 2^5 = 32 possible attribute profiles to indicate if a student has mastered attribute *a* (α_{ea} = 1) or has not mastered attribute *a* (α_{ea} = 0). The LCDM applies the mastery status value to the predicted response pattern for the intercept and the main effect for the attribute measured by the item. Incorporating the connection between LCDM and ANOVA presented previously (Templin & Bradshaw, 2013; Templin & Hoffman, 2013), the LCDM item parameters can be viewed as similar to the levels of independent variables in an analysis of variance (ANOVA), providing an intercept, main effects for each attribute and interactions parameters for items that measure two or more attributes. Because the FCSA included items measure one attribute each, the item parameters include an intercept and a main effect for each item. The intercept ($\lambda_{i,0}$) represents the log-odds of a student correctly answering an item without having mastered the attribute being measured by the item, while the main effect value represents the increase in log-odds of a correct response given mastery of the measured attribute. For example, item 3 is measuring attribute 1, consequently, there are two item parameters noted in Table 12: the intercept ($\lambda_{i,0}$) and the main effect for attribute 1 ($\lambda_{i,1}$). There are no other values for the remaining four main effects because item 3 is not measuring attribute 2, attribute 3, attribute 4 or attribute 5. As seen in Table 3, these values and the mastery status of the student become the terms within the exponent of the item response function (equation1), resulting in item thresholds (τ_{ic}) (Templin & Hoffman, 2013). As there are 32 possible mastery profiles, or latent classes, and 20 items on the FCSA, there are 640 possible τ_{ic} item thresholds (Table 13).

Table 13. LCDM Model Formulas and Thresholds

	Item	*1*	*2*	*3*	*4*	*5*	*6*	*7*
Class	Pattern							
α_{e1}	(0,0,0,0,0)	$\lambda_{1,0}$	$\lambda_{2,0}$	$\lambda_{3,0}$	$\lambda_{4,0}$	$\lambda_{5,0}$	$\lambda_{6,0}$	$\lambda_{7,0}$
α_{e2}	(0,0,0,0,1)	$\lambda_{1,0}$	$\lambda_{2,0}$	$\lambda_{3,0}$	$\lambda_{4,0}$	$\lambda_{5,0}$	$\lambda_{6,0}$	$\lambda_{7,0}$
α_{e3}	(0,0,0,1,0)	$\lambda_{1,0}$	$\lambda_{2,0}$	$\lambda_{3,0}$	$\lambda_{4,0}$	$\lambda_{5,0}$	$\lambda_{6,0}$	$\lambda_{7,0}$
α_{e4}	(0,0,0,1,1)	$\lambda_{1,0}$	$\lambda_{2,0}$	$\lambda_{3,0}$	$\lambda_{4,0}$	$\lambda_{5,0}$	$\lambda_{6,0}$	$\lambda_{7,0}$
α_{e5}	(0,0,1,0,0)	$\lambda_{1,0}$	$\lambda_{2,0}$	$\lambda_{3,0}$	$\lambda_{4,0}$	$\lambda_{5,0}$	$\lambda_{6,0}$	$\lambda_{7,0}$
α_{e6}	(0,0,1,0,1)	$\lambda_{1,0}$	$\lambda_{2,0}$	$\lambda_{3,0}$	$\lambda_{4,0}$	$\lambda_{5,0}$	$\lambda_{6,0}$	$\lambda_{7,0}$
α_{e7}	(0,0,1,1,0)	$\lambda_{1,0}$	$\lambda_{2,0}$	$\lambda_{3,0}$	$\lambda_{4,0}$	$\lambda_{5,0}$	$\lambda_{6,0}$	$\lambda_{7,0}$
α_{e8}	(0,0,1,1,1)	$\lambda_{1,0}$	$\lambda_{2,0}$	$\lambda_{3,0}$	$\lambda_{4,0}$	$\lambda_{5,0}$	$\lambda_{6,0}$	$\lambda_{7,0}$
α_{e9}	(0,1,0,0,0)	$\lambda_{1,0}$	$\lambda_{2,0}$	$\lambda_{3,0}$	$\lambda_{4,0}$	$\lambda_{5,0} + \lambda_{5,1,(2)}$	$\lambda_{6,0} + \lambda_{6,1,(2)}$	$\lambda_{7,0} + \lambda_{7,1,(2)}$
α_{e10}	(0,1,0,0,1)	$\lambda_{1,0}$	$\lambda_{2,0}$	$\lambda_{3,0}$	$\lambda_{4,0}$	$\lambda_{5,0} + \lambda_{5,1,(2)}$	$\lambda_{6,0} + \lambda_{6,1,(2)}$	$\lambda_{7,0} + \lambda_{7,1,(2)}$
α_{e11}	(0,1,0,1,0)	$\lambda_{1,0}$	$\lambda_{2,0}$	$\lambda_{3,0}$	$\lambda_{4,0}$	$\lambda_{5,0} + \lambda_{5,1,(2)}$	$\lambda_{6,0} + \lambda_{6,1,(2)}$	$\lambda_{7,0} + \lambda_{7,1,(2)}$
α_{e12}	(0,1,0,1,1)	$\lambda_{1,0}$	$\lambda_{2,0}$	$\lambda_{3,0}$	$\lambda_{4,0}$	$\lambda_{5,0} + \lambda_{5,1,(2)}$	$\lambda_{6,0} + \lambda_{6,1,(2)}$	$\lambda_{7,0} + \lambda_{7,1,(2)}$
α_{e13}	(0,1,1,0,0)	$\lambda_{1,0}$	$\lambda_{2,0}$	$\lambda_{3,0}$	$\lambda_{4,0}$	$\lambda_{5,0} + \lambda_{5,1,(2)}$	$\lambda_{6,0} + \lambda_{6,1,(2)}$	$\lambda_{7,0} + \lambda_{7,1,(2)}$
α_{e14}	(0,1,1,0,1)	$\lambda_{1,0}$	$\lambda_{2,0}$	$\lambda_{3,0}$	$\lambda_{4,0}$	$\lambda_{5,0} + \lambda_{5,1,(2)}$	$\lambda_{6,0} + \lambda_{6,1,(2)}$	$\lambda_{7,0} + \lambda_{7,1,(2)}$
α_{e15}	(0,1,1,1,0)	$\lambda_{1,0}$	$\lambda_{2,0}$	$\lambda_{3,0}$	$\lambda_{4,0}$	$\lambda_{5,0} + \lambda_{5,1,(2)}$	$\lambda_{6,0} + \lambda_{6,1,(2)}$	$\lambda_{7,0} + \lambda_{7,1,(2)}$
α_{e16}	(0,1,1,1,1)	$\lambda_{1,0}$	$\lambda_{2,0}$	$\lambda_{3,0}$	$\lambda_{4,0}$	$\lambda_{5,0} + \lambda_{5,1,(2)}$	$\lambda_{6,0} + \lambda_{6,1,(2)}$	$\lambda_{7,0} + \lambda_{7,1,(2)}$
α_{e17}	(1,0,0,0,0)	$\lambda_{1,0} + \lambda_{1,1,(1)}$	$\lambda_{1,0} + \lambda_{1,1,(1)}$	$\lambda_{3,0} + \lambda_{3,1,(1)}$	$\lambda_{4,0} + \lambda_{4,1,(1)}$	$\lambda_{5,0}$	$\lambda_{6,0}$	$\lambda_{7,0}$
α_{e18}	(1,0,0,0,1)	$\lambda_{1,0} + \lambda_{1,1,(1)}$	$\lambda_{2,0} + \lambda_{2,1,(1)}$	$\lambda_{3,0} + \lambda_{3,1,(1)}$	$\lambda_{4,0} + \lambda_{4,1,(1)}$	$\lambda_{5,0}$	$\lambda_{6,0}$	$\lambda_{7,0}$
α_{e19}	(1,0,0,1,0)	$\lambda_{1,0} + \lambda_{1,1,(1)}$	$\lambda_{2,0} + \lambda_{2,1,(1)}$	$\lambda_{3,0} + \lambda_{3,1,(1)}$	$\lambda_{4,0} + \lambda_{4,1,(1)}$	$\lambda_{5,0}$	$\lambda_{6,0}$	$\lambda_{7,0}$
α_{e20}	(1,0,0,1,1)	$\lambda_{1,0} + \lambda_{1,1,(1)}$	$\lambda_{2,0} + \lambda_{2,1,(1)}$	$\lambda_{3,0} + \lambda_{3,1,(1)}$	$\lambda_{4,0} + \lambda_{4,1,(1)}$	$\lambda_{5,0}$	$\lambda_{6,0}$	$\lambda_{7,0}$
α_{e21}	(1,0,1,0,0)	$\lambda_{1,0} + \lambda_{1,1,(1)}$	$\lambda_{2,0} + \lambda_{2,1,(1)}$	$\lambda_{3,0} + \lambda_{3,1,(1)}$	$\lambda_{4,0} + \lambda_{4,1,(1)}$	$\lambda_{5,0}$	$\lambda_{6,0}$	$\lambda_{7,0}$
α_{e22}	(1,0,1,0,1)	$\lambda_{1,0} + \lambda_{1,1,(1)}$	$\lambda_{2,0} + \lambda_{2,1,(1)}$	$\lambda_{3,0} + \lambda_{3,1,(1)}$	$\lambda_{4,0} + \lambda_{4,1,(1)}$	$\lambda_{5,0}$	$\lambda_{6,0}$	$\lambda_{7,0}$
α_{e23}	(1,0,1,1,0)	$\lambda_{1,0} + \lambda_{1,1,(1)}$	$\lambda_{2,0} + \lambda_{2,1,(1)}$	$\lambda_{3,0} + \lambda_{3,1,(1)}$	$\lambda_{4,0} + \lambda_{4,1,(1)}$	$\lambda_{5,0}$	$\lambda_{6,0}$	$\lambda_{7,0}$
α_{e24}	(1,0,1,1,1)	$\lambda_{1,0} + \lambda_{1,1,(1)}$	$\lambda_{2,0} + \lambda_{2,1,(1)}$	$\lambda_{3,0} + \lambda_{3,1,(1)}$	$\lambda_{4,0} + \lambda_{4,1,(1)}$	$\lambda_{5,0}$	$\lambda_{6,0}$	$\lambda_{7,0}$
α_{e25}	(1,1,0,0,0)	$\lambda_{1,0} + \lambda_{1,1,(1)}$	$\lambda_{2,0} + \lambda_{2,1,(1)}$	$\lambda_{3,0} + \lambda_{3,1,(1)}$	$\lambda_{4,0} + \lambda_{4,1,(1)}$	$\lambda_{5,0} + \lambda_{5,1,(2)}$	$\lambda_{6,0} + \lambda_{6,1,(2)}$	$\lambda_{7,0} + \lambda_{7,1,(2)}$
α_{e26}	(1,1,0,0,1)	$\lambda_{1,0} + \lambda_{1,1,(1)}$	$\lambda_{2,0} + \lambda_{2,1,(1)}$	$\lambda_{3,0} + \lambda_{3,1,(1)}$	$\lambda_{4,0} + \lambda_{4,1,(1)}$	$\lambda_{5,0} + \lambda_{5,1,(2)}$	$\lambda_{6,0} + \lambda_{6,1,(2)}$	$\lambda_{7,0} + \lambda_{7,1,(2)}$
α_{e27}	(1,1,0,1,0)	$\lambda_{1,0} + \lambda_{1,1,(1)}$	$\lambda_{2,0} + \lambda_{2,1,(1)}$	$\lambda_{3,0} + \lambda_{3,1,(1)}$	$\lambda_{4,0} + \lambda_{4,1,(1)}$	$\lambda_{5,0} + \lambda_{5,1,(2)}$	$\lambda_{6,0} + \lambda_{6,1,(2)}$	$\lambda_{7,0} + \lambda_{7,1,(2)}$
α_{e28}	(1,1,0,1,1)	$\lambda_{1,0} + \lambda_{1,1,(1)}$	$\lambda_{2,0} + \lambda_{2,1,(1)}$	$\lambda_{3,0} + \lambda_{3,1,(1)}$	$\lambda_{4,0} + \lambda_{4,1,(1)}$	$\lambda_{5,0} + \lambda_{5,1,(2)}$	$\lambda_{6,0} + \lambda_{6,1,(2)}$	$\lambda_{7,0} + \lambda_{7,1,(2)}$
α_{e29}	(1,1,1,0,0)	$\lambda_{1,0} + \lambda_{1,1,(1)}$	$\lambda_{2,0} + \lambda_{2,1,(1)}$	$\lambda_{3,0} + \lambda_{3,1,(1)}$	$\lambda_{4,0} + \lambda_{4,1,(1)}$	$\lambda_{5,0} + \lambda_{5,1,(2)}$	$\lambda_{6,0} + \lambda_{6,1,(2)}$	$\lambda_{7,0} + \lambda_{7,1,(2)}$
α_{e30}	(1,1,1,0,1)	$\lambda_{1,0} + \lambda_{1,1,(1)}$	$\lambda_{2,0} + \lambda_{2,1,(1)}$	$\lambda_{3,0} + \lambda_{3,1,(1)}$	$\lambda_{4,0} + \lambda_{4,1,(1)}$	$\lambda_{5,0} + \lambda_{5,1,(2)}$	$\lambda_{6,0} + \lambda_{6,1,(2)}$	$\lambda_{7,0} + \lambda_{7,1,(2)}$
α_{e31}	(1,1,1,1,0)	$\lambda_{1,0} + \lambda_{1,1,(1)}$	$\lambda_{2,0} + \lambda_{2,1,(1)}$	$\lambda_{3,0} + \lambda_{3,1,(1)}$	$\lambda_{4,0} + \lambda_{4,1,(1)}$	$\lambda_{5,0} + \lambda_{5,1,(2)}$	$\lambda_{6,0} + \lambda_{6,1,(2)}$	$\lambda_{7,0} + \lambda_{7,1,(2)}$
α_{e32}	(1,1,1,1,1)	$\lambda_{1,0} + \lambda_{1,1,(1)}$	$\lambda_{2,0} + \lambda_{2,1,(1)}$	$\lambda_{3,0} + \lambda_{3,1,(1)}$	$\lambda_{4,0} + \lambda_{4,1,(1)}$	$\lambda_{5,0} + \lambda_{5,1,(2)}$	$\lambda_{6,0} + \lambda_{6,1,(2)}$	$\lambda_{7,0} + \lambda_{7,1,(2)}$

	Item	*15*	*16*	*17*	*19*	*20*
Class	Pattern					
α_{e1}	(0,0,0,0,0)	$\lambda_{15,0}$	$\lambda_{16,0}$	$\lambda_{17,0}$	$\lambda_{19,0}$	$\lambda_{20,0}$
α_{e2}	(0,0,0,0,1)	$\lambda_{15,0}$	$\lambda_{16,0}$	$\lambda_{17,0}+\lambda_{17,1,(5)}$	$\lambda_{19,0}+\lambda_{19,1,(5)}$	$\lambda_{20,0}+\lambda_{20,1,(5)}$
α_{e3}	(0,0,0,1,0)	$\lambda_{15,0}+\lambda_{15,1,(4)}$	$\lambda_{16,0}+\lambda_{16,1,(4)}$	$\lambda_{17,0}$	$\lambda_{19,0}$	$\lambda_{20,0}$
α_{e4}	(0,0,0,1,1)	$\lambda_{15,0}+\lambda_{15,1,(4)}$	$\lambda_{16,0}+\lambda_{16,1,(4)}$	$\lambda_{17,0}+\lambda_{17,1,(5)}$	$\lambda_{19,0}+\lambda_{19,1,(5)}$	$\lambda_{20,0}+\lambda_{20,1,(5)}$
α_{e5}	(0,0,1,0,0)	$\lambda_{15,0}$	$\lambda_{16,0}$	$\lambda_{17,0}$	$\lambda_{19,0}$	$\lambda_{20,0}$
α_{e6}	(0,0,1,0,1)	$\lambda_{15,0}$	$\lambda_{16,0}$	$\lambda_{17,0}+\lambda_{17,1,(5)}$	$\lambda_{19,0}+\lambda_{19,1,(5)}$	$\lambda_{20,0}+\lambda_{20,1,(5)}$
α_{e7}	(0,0,1,1,0)	$\lambda_{15,0}+\lambda_{15,1,(4)}$	$\lambda_{16,0}+\lambda_{16,1,(4)}$	$\lambda_{17,0}$	$\lambda_{19,0}$	$\lambda_{20,0}$
α_{e8}	(0,0,1,1,1)	$\lambda_{15,0}+\lambda_{15,1,(4)}$	$\lambda_{16,0}+\lambda_{16,1,(4)}$	$\lambda_{17,0}+\lambda_{17,1,(5)}$	$\lambda_{19,0}+\lambda_{19,1,(5)}$	$\lambda_{20,0}+\lambda_{20,1,(5)}$
α_{e9}	(0,1,0,0,0)	$\lambda_{15,0}$	$\lambda_{16,0}$	$\lambda_{17,0}$	$\lambda_{19,0}$	$\lambda_{20,0}$
α_{e10}	(0,1,0,0,1)	$\lambda_{15,0}$	$\lambda_{16,0}$	$\lambda_{17,0}+\lambda_{17,1,(5)}$	$\lambda_{19,0}+\lambda_{19,1,(5)}$	$\lambda_{20,0}+\lambda_{20,1,(5)}$
α_{e11}	(0,1,0,1,0)	$\lambda_{15,0}+\lambda_{15,1,(4)}$	$\lambda_{16,0}+\lambda_{16,1,(4)}$	$\lambda_{17,0}$	$\lambda_{19,0}$	$\lambda_{20,0}$
α_{e12}	(0,1,0,1,1)	$\lambda_{15,0}+\lambda_{15,1,(4)}$	$\lambda_{16,0}+\lambda_{16,1,(4)}$	$\lambda_{17,0}+\lambda_{17,1,(5)}$	$\lambda_{19,0}+\lambda_{19,1,(5)}$	$\lambda_{20,0}+\lambda_{20,1,(5)}$
α_{e13}	(0,1,1,0,0)	$\lambda_{15,0}$	$\lambda_{16,0}$	$\lambda_{17,0}$	$\lambda_{19,0}$	$\lambda_{20,0}$
α_{e14}	(0,1,1,0,1)	$\lambda_{15,0}$	$\lambda_{16,0}$	$\lambda_{17,0}+\lambda_{17,1,(5)}$	$\lambda_{19,0}+\lambda_{19,1,(5)}$	$\lambda_{20,0}+\lambda_{20,1,(5)}$
α_{e15}	(0,1,1,1,0)	$\lambda_{15,0}+\lambda_{15,1,(4)}$	$\lambda_{16,0}+\lambda_{16,1,(4)}$	$\lambda_{17,0}$	$\lambda_{19,0}$	$\lambda_{20,0}$
α_{e16}	(0,1,1,1,1)	$\lambda_{15,0}+\lambda_{15,1,(4)}$	$\lambda_{16,0}+\lambda_{16,1,(4)}$	$\lambda_{17,0}+\lambda_{17,1,(5)}$	$\lambda_{19,0}+\lambda_{19,1,(5)}$	$\lambda_{20,0}+\lambda_{20,1,(5)}$
α_{e17}	(1,0,0,0,0)	$\lambda_{15,0}$	$\lambda_{16,0}$	$\lambda_{17,0}$	$\lambda_{19,0}$	$\lambda_{20,0}$
α_{e18}	(1,0,0,0,1)	$\lambda_{15,0}$	$\lambda_{16,0}$	$\lambda_{17,0}+\lambda_{17,1,(5)}$	$\lambda_{19,0}+\lambda_{19,1,(5)}$	$\lambda_{20,0}+\lambda_{20,1,(5)}$
α_{e19}	(1,0,0,1,0)	$\lambda_{15,0}+\lambda_{15,1,(4)}$	$\lambda_{16,0}+\lambda_{16,1,(4)}$	$\lambda_{17,0}$	$\lambda_{19,0}$	$\lambda_{20,0}$
α_{e20}	(1,0,0,1,1)	$\lambda_{15,0}+\lambda_{15,1,(4)}$	$\lambda_{16,0}+\lambda_{16,1,(4)}$	$\lambda_{17,0}+\lambda_{17,1,(5)}$	$\lambda_{19,0}+\lambda_{19,1,(5)}$	$\lambda_{20,0}+\lambda_{20,1,(5)}$
α_{e21}	(1,0,1,0,0)	$\lambda_{15,0}$	$\lambda_{16,0}$	$\lambda_{17,0}$	$\lambda_{19,0}$	$\lambda_{20,0}$
α_{e22}	(1,0,1,0,1)	$\lambda_{15,0}$	$\lambda_{16,0}$	$\lambda_{17,0}+\lambda_{17,1,(5)}$	$\lambda_{19,0}+\lambda_{19,1,(5)}$	$\lambda_{20,0}+\lambda_{20,1,(5)}$
α_{e23}	(1,0,1,1,0)	$\lambda_{15,0}+\lambda_{15,1,(4)}$	$\lambda_{16,0}+\lambda_{16,1,(4)}$	$\lambda_{17,0}$	$\lambda_{19,0}$	$\lambda_{20,0}$
α_{e24}	(1,0,1,1,1)	$\lambda_{15,0}+\lambda_{15,1,(4)}$	$\lambda_{16,0}+\lambda_{16,1,(4)}$	$\lambda_{17,0}+\lambda_{17,1,(5)}$	$\lambda_{19,0}+\lambda_{19,1,(5)}$	$\lambda_{20,0}+\lambda_{20,1,(5)}$
α_{e25}	(1,1,0,0,0)	$\lambda_{15,0}$	$\lambda_{16,0}$	$\lambda_{17,0}$	$\lambda_{19,0}$	$\lambda_{20,0}$
α_{e26}	(1,1,0,0,1)	$\lambda_{15,0}$	$\lambda_{16,0}$	$\lambda_{17,0}+\lambda_{17,1,(5)}$	$\lambda_{19,0}+\lambda_{19,1,(5)}$	$\lambda_{20,0}+\lambda_{20,1,(5)}$
α_{e27}	(1,1,0,1,0)	$\lambda_{15,0}+\lambda_{15,1,(4)}$	$\lambda_{16,0}+\lambda_{16,1,(4)}$	$\lambda_{17,0}$	$\lambda_{19,0}$	$\lambda_{20,0}$
α_{e28}	(1,1,0,1,1)	$\lambda_{15,0}+\lambda_{15,1,(4)}$	$\lambda_{16,0}+\lambda_{16,1,(4)}$	$\lambda_{17,0}+\lambda_{17,1,(5)}$	$\lambda_{19,0}+\lambda_{19,1,(5)}$	$\lambda_{20,0}+\lambda_{20,1,(5)}$
α_{e29}	(1,1,1,0,0)	$\lambda_{15,0}$	$\lambda_{16,0}$	$\lambda_{17,0}$	$\lambda_{19,0}$	$\lambda_{20,0}$
α_{e30}	(1,1,1,0,1)	$\lambda_{15,0}$	$\lambda_{16,0}$	$\lambda_{17,0}+\lambda_{17,1,(5)}$	$\lambda_{19,0}+\lambda_{19,1,(5)}$	$\lambda_{20,0}+\lambda_{20,1,(5)}$
α_{e31}	(1,1,1,1,0)	$\lambda_{15,0}+\lambda_{15,1,(4)}$	$\lambda_{16,0}+\lambda_{16,1,(4)}$	$\lambda_{17,0}$	$\lambda_{19,0}$	$\lambda_{20,0}$
α_{e32}	(1,1,1,1,1)	$\lambda_{15,0}+\lambda_{15,1,(4)}$	$\lambda_{16,0}+\lambda_{16,1,(4)}$	$\lambda_{17,0}+\lambda_{17,1,(5)}$	$\lambda_{19,0}+\lambda_{19,1,(5)}$	$\lambda_{20,0}+\lambda_{20,1,(5)}$

	Item	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Class	Pattern										
α_{e1}	(0,0,0,0,0)	$\tau_{1,1}$	$\tau_{2,1}$	$\tau_{3,1}$	$\tau_{4,1}$	$\tau_{5,1}$	$\tau_{6,1}$	$\tau_{7,1}$	$\tau_{8,1}$	$\tau_{9,1}$	$\tau_{10,1}$
α_{e2}	(0,0,0,0,1)	$\tau_{1,2}$	$\tau_{2,2}$	$\tau_{3,2}$	$\tau_{4,2}$	$\tau_{5,2}$	$\tau_{6,2}$	$\tau_{7,2}$	$\tau_{8,3}$	$\tau_{9,2}$	$\tau_{10,2}$
α_{e3}	(0,0,0,1,0)	$\tau_{1,3}$	$\tau_{2,3}$	$\tau_{3,3}$	$\tau_{4,3}$	$\tau_{5,3}$	$\tau_{6,3}$	$\tau_{7,3}$	$\tau_{8,3}$	$\tau_{9,3}$	$\tau_{10,3}$
α_{e4}	(0,0,0,1,1)	$\tau_{1,4}$	$\tau_{2,4}$	$\tau_{3,4}$	$\tau_{4,4}$	$\tau_{5,4}$	$\tau_{6,4}$	$\tau_{7,4}$	$\tau_{8,4}$	$\tau_{9,4}$	$\tau_{10,4}$
α_{e5}	(0,0,1,0,0)	$\tau_{1,5}$	$\tau_{2,5}$	$\tau_{3,5}$	$\tau_{4,5}$	$\tau_{5,5}$	$\tau_{6,5}$	$\tau_{7,5}$	$\tau_{8,5}$	$\tau_{9,5}$	$\tau_{10,5}$
α_{e6}	(0,0,1,0,1)	$\tau_{1,6}$	$\tau_{2,6}$	$\tau_{3,6}$	$\tau_{4,6}$	$\tau_{5,6}$	$\tau_{6,6}$	$\tau_{7,6}$	$\tau_{8,6}$	$\tau_{9,6}$	$\tau_{10,6}$
α_{e7}	(0,0,1,1,0)	$\tau_{1,7}$	$\tau_{2,7}$	$\tau_{3,7}$	$\tau_{4,7}$	$\tau_{5,7}$	$\tau_{6,7}$	$\tau_{7,7}$	$\tau_{8,7}$	$\tau_{9,7}$	$\tau_{10,7}$
α_{e8}	(0,0,1,1,1)	$\tau_{1,8}$	$\tau_{2,8}$	$\tau_{3,8}$	$\tau_{4,8}$	$\tau_{5,8}$	$\tau_{6,8}$	$\tau_{7,8}$	$\tau_{8,8}$	$\tau_{9,8}$	$\tau_{10,8}$
α_{e9}	(0,1,0,0,0)	$\tau_{1,9}$	$\tau_{2,9}$	$\tau_{3,9}$	$\tau_{4,9}$	$\tau_{5,9}$	$\tau_{6,9}$	$\tau_{7,9}$	$\tau_{8,9}$	$\tau_{9,9}$	$\tau_{10,9}$
α_{e10}	(0,1,0,0,1)	$\tau_{1,10}$	$\tau_{2,10}$	$\tau_{3,10}$	$\tau_{4,10}$	$\tau_{5,10}$	$\tau_{6,10}$	$\tau_{7,10}$	$\tau_{8,10}$	$\tau_{9,10}$	$\tau_{10,10}$
α_{e11}	(0,1,0,1,0)	$\tau_{1,11}$	$\tau_{2,11}$	$\tau_{3,11}$	$\tau_{4,11}$	$\tau_{5,11}$	$\tau_{6,11}$	$\tau_{7,11}$	$\tau_{8,11}$	$\tau_{9,11}$	$\tau_{10,11}$
α_{e12}	(0,1,0,1,1)	$\tau_{1,12}$	$\tau_{2,12}$	$\tau_{3,12}$	$\tau_{4,12}$	$\tau_{5,12}$	$\tau_{6,12}$	$\tau_{7,12}$	$\tau_{8,12}$	$\tau_{9,12}$	$\tau_{10,12}$
α_{e13}	(0,1,1,0,0)	$\tau_{1,13}$	$\tau_{2,13}$	$\tau_{3,13}$	$\tau_{4,13}$	$\tau_{5,13}$	$\tau_{6,13}$	$\tau_{7,13}$	$\tau_{8,13}$	$\tau_{9,13}$	$\tau_{10,13}$
α_{e14}	(0,1,1,0,1)	$\tau_{1,14}$	$\tau_{2,14}$	$\tau_{3,14}$	$\tau_{4,14}$	$\tau_{5,14}$	$\tau_{6,14}$	$\tau_{7,14}$	$\tau_{8,14}$	$\tau_{9,14}$	$\tau_{10,14}$
α_{e15}	(0,1,1,1,0)	$\tau_{1,15}$	$\tau_{2,15}$	$\tau_{3,15}$	$\tau_{4,15}$	$\tau_{5,15}$	$\tau_{6,15}$	$\tau_{7,15}$	$\tau_{8,15}$	$\tau_{9,15}$	$\tau_{10,15}$
α_{e16}	(0,1,1,1,1)	$\tau_{1,16}$	$\tau_{2,16}$	$\tau_{3,16}$	$\tau_{4,16}$	$\tau_{5,16}$	$\tau_{6,16}$	$\tau_{7,16}$	$\tau_{8,16}$	$\tau_{9,16}$	$\tau_{10,16}$
α_{e17}	(1,0,0,0,0)	$\tau_{1,17}$	$\tau_{2,17}$	$\tau_{3,17}$	$\tau_{4,17}$	$\tau_{5,17}$	$\tau_{6,17}$	$\tau_{7,17}$	$\tau_{8,17}$	$\tau_{9,17}$	$\tau_{10,17}$
α_{e18}	(1,0,0,0,1)	$\tau_{1,18}$	$\tau_{2,18}$	$\tau_{3,18}$	$\tau_{4,18}$	$\tau_{5,18}$	$\tau_{6,18}$	$\tau_{7,18}$	$\tau_{8,18}$	$\tau_{9,18}$	$\tau_{10,18}$
α_{e19}	(1,0,0,1,0)	$\tau_{1,19}$	$\tau_{2,19}$	$\tau_{3,19}$	$\tau_{4,19}$	$\tau_{5,19}$	$\tau_{6,19}$	$\tau_{7,19}$	$\tau_{8,19}$	$\tau_{9,19}$	$\tau_{10,19}$
α_{e20}	(1,0,0,1,1)	$\tau_{1,20}$	$\tau_{2,20}$	$\tau_{3,20}$	$\tau_{4,20}$	$\tau_{5,20}$	$\tau_{6,20}$	$\tau_{7,20}$	$\tau_{8,20}$	$\tau_{9,20}$	$\tau_{10,20}$
α_{e21}	(1,0,1,0,0)	$\tau_{1,21}$	$\tau_{2,21}$	$\tau_{3,21}$	$\tau_{4,21}$	$\tau_{5,21}$	$\tau_{6,21}$	$\tau_{7,21}$	$\tau_{8,21}$	$\tau_{9,21}$	$\tau_{10,21}$
α_{e22}	(1,0,1,0,1)	$\tau_{1,22}$	$\tau_{2,22}$	$\tau_{3,22}$	$\tau_{4,22}$	$\tau_{5,22}$	$\tau_{6,22}$	$\tau_{7,22}$	$\tau_{8,22}$	$\tau_{9,22}$	$\tau_{10,22}$
α_{e23}	(1,0,1,1,0)	$\tau_{1,23}$	$\tau_{2,23}$	$\tau_{3,23}$	$\tau_{4,23}$	$\tau_{5,23}$	$\tau_{6,23}$	$\tau_{7,23}$	$\tau_{8,23}$	$\tau_{9,23}$	$\tau_{10,23}$
α_{e24}	(1,0,1,1,1)	$\tau_{1,24}$	$\tau_{2,24}$	$\tau_{3,24}$	$\tau_{4,24}$	$\tau_{5,24}$	$\tau_{6,24}$	$\tau_{7,24}$	$\tau_{8,24}$	$\tau_{9,24}$	$\tau_{10,24}$
α_{e25}	(1,1,0,0,0)	$\tau_{1,25}$	$\tau_{2,25}$	$\tau_{3,25}$	$\tau_{4,25}$	$\tau_{5,25}$	$\tau_{6,25}$	$\tau_{7,25}$	$\tau_{8,25}$	$\tau_{9,25}$	$\tau_{10,25}$
α_{e26}	(1,1,0,0,1)	$\tau_{1,26}$	$\tau_{2,26}$	$\tau_{3,26}$	$\tau_{4,26}$	$\tau_{5,26}$	$\tau_{6,26}$	$\tau_{7,26}$	$\tau_{8,26}$	$\tau_{9,26}$	$\tau_{10,26}$
α_{e27}	(1,1,0,1,0)	$\tau_{1,27}$	$\tau_{2,27}$	$\tau_{3,27}$	$\tau_{4,27}$	$\tau_{5,27}$	$\tau_{6,27}$	$\tau_{7,27}$	$\tau_{8,27}$	$\tau_{9,27}$	$\tau_{10,27}$
α_{e28}	(1,1,0,1,1)	$\tau_{1,28}$	$\tau_{2,28}$	$\tau_{3,28}$	$\tau_{4,28}$	$\tau_{5,28}$	$\tau_{6,28}$	$\tau_{7,28}$	$\tau_{8,28}$	$\tau_{9,28}$	$\tau_{10,28}$
α_{e29}	(1,1,1,0,0)	$\tau_{1,29}$	$\tau_{2,29}$	$\tau_{3,29}$	$\tau_{4,29}$	$\tau_{5,29}$	$\tau_{6,29}$	$\tau_{7,29}$	$\tau_{8,29}$	$\tau_{9,29}$	$\tau_{10,29}$
α_{e30}	(1,1,1,0,1)	$\tau_{1,30}$	$\tau_{2,30}$	$\tau_{3,30}$	$\tau_{4,30}$	$\tau_{5,30}$	$\tau_{6,30}$	$\tau_{7,30}$	$\tau_{8,30}$	$\tau_{9,30}$	$\tau_{10,30}$
α_{e31}	(1,1,1,1,0)	$\tau_{1,31}$	$\tau_{2,31}$	$\tau_{3,31}$	$\tau_{4,31}$	$\tau_{5,31}$	$\tau_{6,31}$	$\tau_{7,31}$	$\tau_{8,31}$	$\tau_{9,31}$	$\tau_{10,31}$
α_{e32}	(1,1,1,1,1)	$\tau_{1,32}$	$\tau_{2,32}$	$\tau_{3,21}$	$\tau_{4,32}$	$\tau_{5,32}$	$\tau_{6,32}$	$\tau_{7,32}$	$\tau_{8,32}$	$\tau_{9,32}$	$\tau_{10,32}$

	Item	11	12	13	14	15	16	17	19	20
Class	Pattern									
α_{e1}	(0,0,0,0,0)	$\tau_{11,1}$	$\tau_{12,1}$	$\tau_{13,1}$	$\tau_{14,1}$	$\tau_{15,1}$	$\tau_{16,1}$	$\tau_{17,1}$	$\tau_{19,1}$	$\tau_{20,1}$
α_{e2}	(0,0,0,0,1)	$\tau_{11,2}$	$\tau_{12,2}$	$\tau_{13,2}$	$\tau_{14,2}$	$\tau_{15,2}$	$\tau_{16,2}$	$\tau_{17,2}$	$\tau_{19,2}$	$\tau_{20,2}$
α_{e3}	(0,0,0,1,0)	$\tau_{11,3}$	$\tau_{12,3}$	$\tau_{13,3}$	$\tau_{14,3}$	$\tau_{15,3}$	$\tau_{16,3}$	$\tau_{17,3}$	$\tau_{19,3}$	$\tau_{20,3}$
α_{e4}	(0,0,0,1,1)	$\tau_{11,4}$	$\tau_{12,4}$	$\tau_{13,4}$	$\tau_{14,4}$	$\tau_{15,4}$	$\tau_{16,4}$	$\tau_{17,4}$	$\tau_{19,4}$	$\tau_{20,4}$
α_{e5}	(0,0,1,0,0)	$\tau_{11,5}$	$\tau_{12,5}$	$\tau_{13,5}$	$\tau_{14,5}$	$\tau_{15,5}$	$\tau_{16,5}$	$\tau_{17,5}$	$\tau_{19,5}$	$\tau_{20,5}$
α_{e6}	(0,0,1,0,1)	$\tau_{11,6}$	$\tau_{12,6}$	$\tau_{13,6}$	$\tau_{14,6}$	$\tau_{15,6}$	$\tau_{16,6}$	$\tau_{17,6}$	$\tau_{19,6}$	$\tau_{20,6}$
α_{e7}	(0,0,1,1,0)	$\tau_{11,7}$	$\tau_{12,7}$	$\tau_{13,7}$	$\tau_{14,7}$	$\tau_{15,7}$	$\tau_{16,7}$	$\tau_{17,7}$	$\tau_{19,7}$	$\tau_{20,7}$
α_{e8}	(0,0,1,1,1)	$\tau_{11,8}$	$\tau_{12,8}$	$\tau_{13,8}$	$\tau_{14,8}$	$\tau_{15,8}$	$\tau_{16,8}$	$\tau_{17,8}$	$\tau_{19,8}$	$\tau_{20,8}$
α_{e9}	(0,1,0,0,0)	$\tau_{11,9}$	$\tau_{12,9}$	$\tau_{13,9}$	$\tau_{14,9}$	$\tau_{15,9}$	$\tau_{16,9}$	$\tau_{17,9}$	$\tau_{19,9}$	$\tau_{20,9}$
α_{e10}	(0,1,0,0,1)	$\tau_{11,10}$	$\tau_{12,10}$	$\tau_{13,10}$	$\tau_{14,10}$	$\tau_{15,10}$	$\tau_{16,10}$	$\tau_{17,10}$	$\tau_{19,10}$	$\tau_{20,10}$
α_{e11}	(0,1,0,1,0)	$\tau_{11,11}$	$\tau_{12,11}$	$\tau_{13,11}$	$\tau_{14,11}$	$\tau_{15,11}$	$\tau_{16,11}$	$\tau_{17,11}$	$\tau_{19,11}$	$\tau_{20,11}$
α_{e12}	(0,1,0,1,1)	$\tau_{11,12}$	$\tau_{12,12}$	$\tau_{13,12}$	$\tau_{14,12}$	$\tau_{15,12}$	$\tau_{16,12}$	$\tau_{17,12}$	$\tau_{19,12}$	$\tau_{20,12}$
α_{e13}	(0,1,1,0,0)	$\tau_{11,13}$	$\tau_{12,13}$	$\tau_{13,13}$	$\tau_{14,13}$	$\tau_{15,13}$	$\tau_{16,13}$	$\tau_{17,13}$	$\tau_{19,13}$	$\tau_{20,13}$
α_{e14}	(0,1,1,0,1)	$\tau_{11,14}$	$\tau_{12,14}$	$\tau_{13,14}$	$\tau_{14,14}$	$\tau_{15,14}$	$\tau_{16,14}$	$\tau_{17,14}$	$\tau_{19,14}$	$\tau_{20,14}$
α_{e15}	(0,1,1,1,0)	$\tau_{11,15}$	$\tau_{12,15}$	$\tau_{13,15}$	$\tau_{14,15}$	$\tau_{15,15}$	$\tau_{16,15}$	$\tau_{17,15}$	$\tau_{19,15}$	$\tau_{20,15}$
α_{e16}	(0,1,1,1,1)	$\tau_{11,16}$	$\tau_{12,16}$	$\tau_{13,16}$	$\tau_{14,16}$	$\tau_{15,16}$	$\tau_{16,16}$	$\tau_{17,16}$	$\tau_{19,16}$	$\tau_{20,16}$
α_{e17}	(1,0,0,0,0)	$\tau_{11,17}$	$\tau_{12,17}$	$\tau_{13,17}$	$\tau_{14,17}$	$\tau_{15,17}$	$\tau_{16,17}$	$\tau_{17,17}$	$\tau_{19,17}$	$\tau_{20,17}$
α_{e18}	(1,0,0,0,1)	$\tau_{11,18}$	$\tau_{12,18}$	$\tau_{13,18}$	$\tau_{14,18}$	$\tau_{15,18}$	$\tau_{16,18}$	$\tau_{17,18}$	$\tau_{19,18}$	$\tau_{20,18}$
α_{e19}	(1,0,0,1,0)	$\tau_{11,19}$	$\tau_{12,19}$	$\tau_{13,19}$	$\tau_{14,19}$	$\tau_{15,19}$	$\tau_{16,19}$	$\tau_{17,19}$	$\tau_{19,19}$	$\tau_{20,19}$
α_{e20}	(1,0,0,1,1)	$\tau_{11,20}$	$\tau_{12,20}$	$\tau_{13,20}$	$\tau_{14,20}$	$\tau_{15,20}$	$\tau_{16,20}$	$\tau_{17,20}$	$\tau_{19,20}$	$\tau_{20,20}$
α_{e21}	(1,0,1,0,0)	$\tau_{11,21}$	$\tau_{12,21}$	$\tau_{13,21}$	$\tau_{14,21}$	$\tau_{15,21}$	$\tau_{16,21}$	$\tau_{17,21}$	$\tau_{19,21}$	$\tau_{20,21}$
α_{e22}	(1,0,1,0,1)	$\tau_{11,22}$	$\tau_{12,22}$	$\tau_{13,22}$	$\tau_{14,22}$	$\tau_{15,22}$	$\tau_{16,22}$	$\tau_{17,22}$	$\tau_{19,22}$	$\tau_{20,22}$
α_{e23}	(1,0,1,1,0)	$\tau_{11,23}$	$\tau_{12,23}$	$\tau_{13,23}$	$\tau_{14,23}$	$\tau_{15,23}$	$\tau_{16,23}$	$\tau_{17,23}$	$\tau_{19,23}$	$\tau_{20,23}$
α_{e24}	(1,0,1,1,1)	$\tau_{11,24}$	$\tau_{12,24}$	$\tau_{13,24}$	$\tau_{14,24}$	$\tau_{15,24}$	$\tau_{16,24}$	$\tau_{17,24}$	$\tau_{19,24}$	$\tau_{20,24}$
α_{e25}	(1,1,0,0,0)	$\tau_{11,25}$	$\tau_{12,25}$	$\tau_{13,25}$	$\tau_{14,25}$	$\tau_{15,25}$	$\tau_{16,25}$	$\tau_{17,25}$	$\tau_{19,25}$	$\tau_{20,25}$
α_{e26}	(1,1,0,0,1)	$\tau_{11,26}$	$\tau_{12,26}$	$\tau_{13,26}$	$\tau_{14,26}$	$\tau_{15,26}$	$\tau_{16,26}$	$\tau_{17,26}$	$\tau_{19,26}$	$\tau_{20,26}$
α_{e27}	(1,1,0,1,0)	$\tau_{11,27}$	$\tau_{12,27}$	$\tau_{13,27}$	$\tau_{14,27}$	$\tau_{15,27}$	$\tau_{16,27}$	$\tau_{17,27}$	$\tau_{19,27}$	$\tau_{20,27}$
α_{e28}	(1,1,0,1,1)	$\tau_{11,28}$	$\tau_{12,28}$	$\tau_{13,28}$	$\tau_{14,28}$	$\tau_{15,28}$	$\tau_{16,28}$	$\tau_{17,28}$	$\tau_{19,28}$	$\tau_{20,28}$
α_{e29}	(1,1,1,0,0)	$\tau_{11,29}$	$\tau_{12,29}$	$\tau_{13,29}$	$\tau_{14,29}$	$\tau_{15,29}$	$\tau_{16,29}$	$\tau_{17,29}$	$\tau_{19,29}$	$\tau_{20,29}$
α_{e30}	(1,1,1,0,1)	$\tau_{11,30}$	$\tau_{12,30}$	$\tau_{13,30}$	$\tau_{14,30}$	$\tau_{15,30}$	$\tau_{16,30}$	$\tau_{17,30}$	$\tau_{19,30}$	$\tau_{20,30}$
α_{e31}	(1,1,1,1,0)	$\tau_{11,31}$	$\tau_{12,31}$	$\tau_{13,31}$	$\tau_{14,31}$	$\tau_{15,31}$	$\tau_{16,31}$	$\tau_{17,31}$	$\tau_{19,31}$	$\tau_{20,31}$
α_{e32}	(1,1,1,1,1)	$\tau_{11,32}$	$\tau_{12,32}$	$\tau_{13,32}$	$\tau_{14,32}$	$\tau_{15,32}$	$\tau_{16,32}$	$\tau_{17,32}$	$\tau_{19,32}$	$\tau_{20,32}$

The LCDM item parameters and the Q-matrix mapping were used to complete the item response function for each of the 640 item thresholds. For each of the 32 classes, the item response function was calculated based on mastery and non-mastery of each attribute. For example, for attribute 1 each attribute profile had an item response function for mastery of attribute 1 and one for non-mastery of attribute 1. For the attribute profiles in which attribute 1 was mastered, the LCDM item response function was:

$$\tau_{i,c} = \lambda_{i,0} + \lambda_{i,1,(1)}\alpha_{c1} = \lambda_{i,0} + \lambda_{i,1,(1)} \tag{5}$$

For the attribute profiles in which attribute 1 was not mastered, the LCDM item response function was:

$$\tau_{i,c} = \lambda_{i,0} + \lambda_{i,1,(1)}\alpha_{c1} = \lambda_{i,0} + \lambda_{i,1,(1)}(0) = \lambda_{i,0} \tag{6}$$

For example, item 3 measures attribute 1, consequently, each of the class-specific item thresholds (τ_{3c}) for item 3 was defined by equation 2 for the classes where attribute 1 was mastered (α_{e17} - α_{e32}), and equation 3 for the classes where attribute 1 was not mastered (α_{e1} - α_{e16}). The LCDM also places order constraints on the item parameters for the main effects, ensuring that students that have mastered the measured attribute receive a higher probability of a correct response (Templin & Hoffman).

Structural Model Parameters

For each of the mastery profiles, or latent classes, the LCDM provides the estimated number of students that have the profile. Additionally, the structural parameters (υ_c) that represent the proportion of the data set that is classified as a member of each mastery profile are provided as well (Table 14). Following Templin and Hoffman (2013), particular attention is paid to results based on the estimated model, which presents the most likely estimate of each structural parameter.

Table 14. LCDM Estimate of Expected Count and Proportion of Students by Mastery Profile.

Class	Pattern	Expected Count	υ_c
α_{e1}	(0,0,0,0,0)	886.03	0.541
α_{e2}	(0,0,0,0,1)	39.73	0.024
α_{e3}	(0,0,0,1,0)	1.57	0.001
α_{e4}	(0,0,0,1,1)	40.86	0.025
α_{e5}	(0,0,1,0,0)	12.29	0.008
α_{e6}	(0,0,1,0,1)	0.00	0.000
α_{e7}	(0,0,1,1,0)	12.67	0.008
α_{e8}	(0,0,1,1,1)	27.74	0.017
α_{e9}	(0,1,0,0,0)	35.54	0.022
α_{e10}	(0,1,0,0,1)	0.00	0.000
α_{e11}	(0,1,0,1,0)	0.00	0.000
α_{e12}	(0,1,0,1,1)	22.70	0.014
α_{e13}	(0,1,1,0,0)	6.99	0.004
α_{e14}	(0,1,1,0,1)	0.00	0.000
α_{e15}	(0,1,1,1,0)	1.05	0.001
α_{e16}	(0,1,1,1,1)	124.36	0.076
α_{e17}	(1,0,0,0,0)	46.18	0.028
α_{e18}	(1,0,0,0,1)	18.36	0.011
α_{e19}	(1,0,0,1,0)	2.79	0.002
α_{e20}	(1,0,0,1,1)	0.00	0.000
α_{e21}	(1,0,1,0,0)	7.80	0.005
α_{e22}	(1,0,1,0,1)	0.00	0.000
α_{e23}	(1,0,1,1,0)	0.00	0.000
α_{e24}	(1,0,1,1,1)	21.51	0.013
α_{e25}	(1,1,0,0,0)	14.64	0.009
α_{e26}	(1,1,0,0,1)	10.87	0.007
α_{e27}	(1,1,0,1,0)	2.60	0.002
α_{e28}	(1,1,0,1,1)	9.82	0.006
α_{e29}	(1,1,1,0,0)	1.96	0.001
α_{e30}	(1,1,1,0,1)	3.55	0.002
α_{e31}	(1,1,1,1,0)	17.80	0.011
α_{e32}	(1,1,1,1,1)	269.61	0.164

Note:

Looking at the mastery profiles, or latent classes, that exhibited the highest proportion of students from the data set, it is obvious that the majority of the students were classified as α_{e1} =

.541, meaning about 54% of the students had not having mastered any of the five attributes. The next largest mastery profile was α_{e32} = .164, indicating that approximately 16% of the students had mastered all of the attributes measured by the FCSA. Interestingly, the next largest profile was α_{e16} = .076, meaning almost 8% of the students that had taken the FCSA were classified as having mastered attributes 2, 3, 4, and 5. This is in direct contrast to the structure of FCSAH, as the hierarchy indicates that attribute 1 must be learned before a student can progress to attribute 2, which is followed by any combination of attributes 3-5. These findings highlight the necessity and value in proceeding with the HDCM analysis to statistically determine if there is in fact a learning hierarchy underlying the structure of the FCSA.

HDCM Results

HDCM can be used to evaluate and test for the presence of an attribute hierarchy by constraining the parameters of the LCDM to zero, allowing a researcher to statistically confirm or falsify a theorized learning hierarchy (Templin and Bradshaw, 2013). As noted in the research, Attribute Hierarchy Method (AHM) does not allow for testing of the hypothesis that a hierarchy exists, relying predominantly on goodness of fit information to summarize model fit (Templin and Bradshaw), consequently, proceeding with a comparison of the LCDM to the HDCM, constrained to align with the hierarchy outlined in the FCSAH, provides that analysis.

Because the FCSA was designed with a specific learning hierarchy, the HDCM was used as a constrained model of the fully crossed LCDM by fixing the redundant parameters created by the LCDM to zero. The structural model is reduced as the parameters that represented attribute profiles that were not possible were removed for the HDCM analysis. Recall, the number of attribute profiles estimated under the LCDM was 2^5, while HDCM estimates only A + 1 attribute profiles; consequently, the number of estimated attribute profiles decreases from 2^5 = 32 possible attribute profiles down to 6 profiles estimated. Under the FCSAH, a student must master attribute 1, then must master attribute 2 before proceeding to any combination of attribute 3, 4, or 5 (Figure 2), consequently, ten of the attribute profiles from the 32 possible profiles seen in the LCDM were retained for the HDCM analysis to align with the FCSAH (Table 15). This highlights that under the FCSAH there are attributes that represent *nested* factors wherein three attributes are dependent on two other attributes, these can then be considered as *nested* within the dependent attributes. Under HDCM the attributes in profile *c* are represented as $\alpha^{*}{}_{c}$ and the set of item parameters need to reflect the nested structure of the attribute profiles, so the matrix portion of item response function changes. However, because each item included in the FCSA measured only one attribute, the item parameters remain the same as those from the LCDM, resulting in an intercept and main effect for each item (Table 16).

Table 15. Attribute Mastery Profiles Retained Under the HDCM

Mastery profile	Attribute 1	Attribute 2	Attribute 3	Attribute 4	Attribute 5
α_{e1}	0	0	0	0	0
α_{e17}	1	0	0	0	0
α_{e25}	1	1	0	0	0
α_{e26}	1	1	0	0	1
α_{e27}	1	1	0	1	0
α_{e28}	1	1	0	1	1
α_{e29}	1	1	1	0	0
α_{e30}	1	1	1	0	1
α_{e31}	1	1	1	1	0
α_{e32}	1	1	1	1	1

Note: mastery = 1; non-mastery = 0

Table 16. FCSA Q-Matrix and HDCM Item Parameter Estimates

i	Skill Measured	Item Difficulty	$\lambda_{i,0}$	$\lambda_{i,1}$	$\lambda_{i,2}$	$\lambda_{i,3}$	$\lambda_{i,4}$	$\lambda_{i,5}$
1	A1	-2.35	3.19(0.30)	1.79(0.33)				
2	A1	-2.95	8.09(14.18)	5.77(14.19)				
3	A1	-2.03	3.50(0.38)	2.23(0.40)				
4	A1	-1.47	3.44(0.48)	2.62(0.28)				
5	A2	-2.06	3.92(0.27)		2.53(0.31)			
6	A2	-1.17	2.72(0.18)		2.08(0.20)			
7	A2	0.00	1.37(0.11)		1.93(0.15)			
8	A2	-1.15	2.34(0.13)		1.75(0.17)			
9	A3	-0.76	1.85(0.10)			1.90(0.15)		
10	A3	-0.82	3.13(0.20)			3.39(0.22)		
11	A3	-0.43	1.66(0.10)			2.35(0.15)		
12	A3	-1.27	3.60(0.22)			3.17(0.24)		
13	A4	-0.03	1.07(0.09)				2.33(0.16)	
14	A4	-0.95	2.70(0.16)				2.76(0.19)	
15	A4	-0.08	1.13(0.09)				2.04(0.14)	
16	A4	-1.51	2.42(0.13)				1.67(0.17)	
17	A5	-1.23	2.38(0.13)					2.12(0.18)
18	A5	-1.20						
19	A5	-0.24	1.49(0.10)					2.41(0.21)
20	A5	2.09	-0.54(0.07)					0.55(0.15)

Note: Item 18 was removed in final analysis.

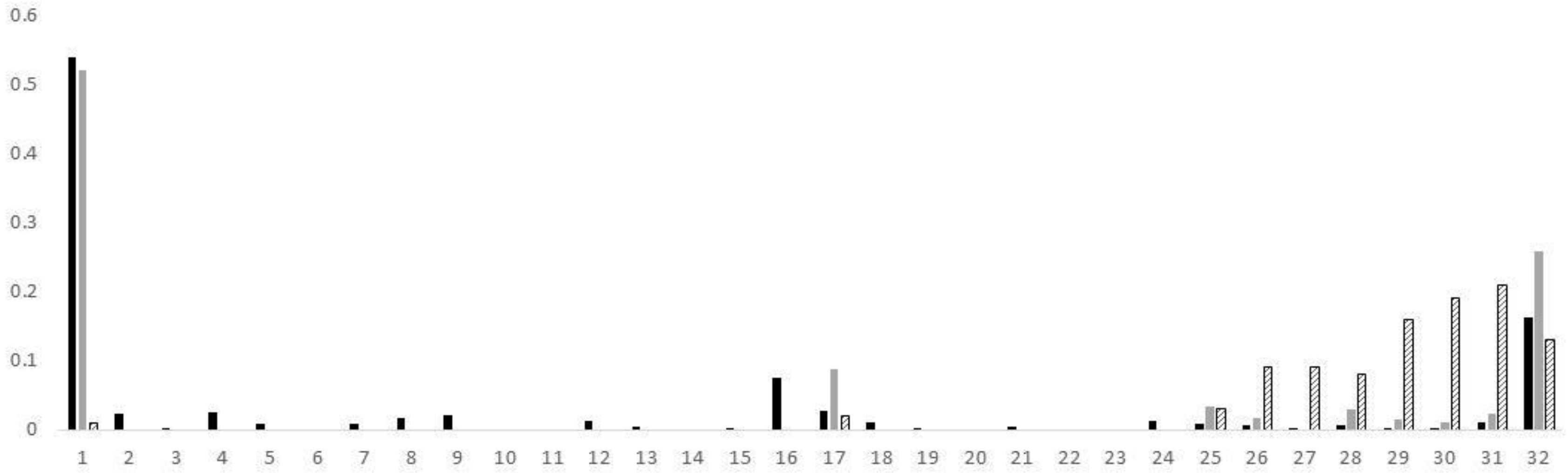
Again, several profiles within the LCDM were constrained to zero in the HDCM model based on the structure of the FCSAH. By removing these profiles, it was expected that the model fit would improve as the model now represented the hierarchy. However, the overall model fit decreased under the HDCM as evidenced by the increase in the overall bivariate chi-square value; from χ^2 = 289.79 of the LCDM analysis to χ^2 = 337.27 in the HDCM analysis. This increase indicates that by constraining the LCDM to align to the FCSAH, the model fit has suffered; using a model constrained to the hierarchy erodes the model fit, there is reason to question the presence of a learning hierarchy.

Structural Model Parameters

As with the LCDM, for each of the mastery profiles the HDCM provides the estimated number of students that have the profile. Additionally, the structural parameters (υ_c) that represent the proportion of the data set that is classified as a member of each mastery profile are provided as well (Table 17). Under the HDCM, mastery classification was forced into ten attribute mastery profiles that match the learning hierarchy; consequently, the model provides expected counts and data set proportions limited to those profiles. Figure 3 outlines the distribution of the structural parameters under the LCDM and HDCM, clearly highlighting that the majority of the students in the data set were classified as having mastered none of the attributes, followed by the profile indicative of having mastered all of the attributes measured in the FCSA. Considering the three profiles with the largest proportions, under the LCDM, approximately 78% of the students were classified as having mastered none of the attributes (54%), all of the attributes (16%), or in the profile representing mastery of attributes 2, 3, 4 and 5 (8%). Under the HDCM, the three profiles receiving the largest proportions of students classified were no attributes mastered (52%), all attributes mastered (26%) and mastering only the first attribute (9%).

Table 17. HDCM Estimate of Expected Count and Proportion of Students by Mastery Profile.

Class	Pattern	Expected Count	υ_c
α_{e1}	(0,0,0,0,0)	855.77	0.522
α_{e17}	(1,0,0,0,0)	144.31	0.088
α_{e25}	(1,1,0,0,0)	56.98	0.035
α_{e26}	(1,1,0,0,1)	27.12	0.017
α_{e27}	(1,1,0,1,0)	0.00	0.000
α_{e28}	(1,1,0,1,1)	49.15	0.030
α_{e29}	(1,1,1,0,0)	26.10	0.016
α_{e30}	(1,1,1,0,1)	20.20	0.012
α_{e31}	(1,1,1,1,0)	37.31	0.023
α_{e32}	(1,1,1,1,1)	422.06	0.258



Profile	1	2	4	5	7	8	9	12	13	15	16	17	18	19	21	24	25	26	27	28	29	30	31	32
LCDM	.54	.02	.03	.01	.01	.02	.02	.01	.001	.001	.08	.03	.01	.001	.01	.01	.01	.01	.001	.01	.00	.001	.01	.16
HDCM	.52	--	--	--	--	--	--	--	--	--	--	.09	--	--	--	--	.04	.02	--	.03	.02	.01	.02	.26
AHM	.01	--	--	--	--	--	--	--	--	--	--	.02	--	--	--	--	.03	.09	.09	.08	.16	.19	.21	.13

Figure 3.

Structural parameter estimates under the LCDM and the HDCM.

Statistical Analysis of Attribute Hierarchy

The central component of applying the HDCM is to statistically evaluate the presence of an attribute hierarchy and analyze the fit of the two models, one simpler and one more complex. By comparing deviances, equivalent to a likelihood ratio test with the number of degrees of freedom for the chi-squared being the difference in the number of parameters in the two models, this statistical analysis is possible (Templin and Bradshaw, 2013). Consequently, to evaluate the model comparison of the LCDM to the nested HDCM, this study used a deviance goodness of fit test for the statistical hypothesis testing of -2 times the difference in model log-likelihood ratio of the reduced model compared to the full model, which was 68.408. This was compared to a Chi-Square distribution with 22 degrees of freedom (the difference in the number of model parameters) using pchisq in R, resulting in a $p < .001$. The null hypothesis is that the model is correctly specified, however, the significance of the deviance statistic indicates there is strong evidence to reject that hypothesis, meaning the reduced model does not fit the data well and there does not appear to be an attribute hierarchy present in the data; consequently, the LCDM is the best fitting model.

Comparison to previous Attribute Hierarchy Method Classification Output

This study was designed to evaluate the ability of the HDCM to identify and/or confirm the presence of a theorized learning attribute hierarchy in K12 education, using the formative FCSA assessment which was developed based on such a hierarchy. The previous AHM analysis of the FCSA (Broaddus, 2011) established ten knowledge states, or mastery profiles, which resulted in the majority of the students within the data set to be classified as having mastered many, if not all, of the attributes of the FCSAH (see Table 9). As seen in Figure 3, the classifications from the AHM are drastically different than what was found with the LCDM and the HDCM, where both models classified the majority of students into profiles indicative of having mastered either none or all of the attributes. These disparities in classification proportions is likely due to the use of a cognitive model that assumes the presence of an attribute hierarchy with a data set that does not support that assumption. Had the hierarchy been present, the HDCM would have exhibited better fit than the LCDM due the parsimony of the model, thereby supporting the use of AHM in analyzing the data; however, based on the output reviewed, this in fact is not the case as the LCDM fit was better.

CHAPTER 5

Discussion

There have been numerous arguments evaluating the presence of learning attributes within education, as the collective understanding is that learning occurs in steps at some points. It is these steps that formulate a hierarchy and require a student to learn one concept or skill prior to being able to complete another; consequently, there is a sequential order to most education processes. The pedagogical theories underlying curriculum creation support sequential learning, therefore, student responses and response patterns should represent the student's current level of understanding and provide insight into the resulting hierarchy (Templin & Bradshaw, 2013).

Formative assessment, and more importantly the effective applications of formative processes, provides educators with the most up-to-date, real-time gauges of student understanding in smaller learning units, that is, clearly defined steps or stages in sequence. One of the most influential formative processes is feedback that, in the most basic forms, identifies for a student what the end goal is, where the student is currently, and what needs to be done to get to the end goal. Lesson plans, classroom activities and assessments work well in defining what the end goal is, be it the standards that are being measured or the act of correctly performing a specified task. By using formative assessment, educators can clearly stipulate the expected outcomes of a learning activity, task or assessment. For example, an educator has created a lesson plan designed to teach students to successfully divide three digit numbers and that lesson plan includes several classroom activities, group discussions and practice/homework pieces. At the end of the lesson plan, the educator gives the class an assessment covering the division of three digit numbers. The feedback received throughout the lesson plan that is most valuable in increasing student understanding reiterates the end goal, while identifying where the student is currently on the “road” to that end goal and, importantly, what the student needs to do next to progress toward that end goal. Consider a map from one state to another; if a driver is aware of the final destination that is one piece of information, however, the driver must also know where they are at any point and particularly what roads and actions are still needed to reach the destination. This can be extended to student understanding. To illuminate the current understanding and the next steps requires the identification of what skills or attributes a student has mastered or has not mastered. Unfortunately, most educational assessments do not provide this discrete level of information as they are not developed with a strong integration of cognitive theory, nor do the assessments use scoring models that would support providing such information. This is not a novel awareness as noted in the previous chapters of this study, and while research and test development are advancing to create such links, it is important to be aware that cognitive diagnostic models can work with typical scoring models to provide student ability information as well as information regarding student understanding gleaned from the items the student answered not only correctly, but incorrectly as well.

To revisit, cognitive diagnostic models (CDM) can be described in the most general terms as models that identify or classify an examinee’s current level understanding based on the examinee’s response patterns in relationship to the attributes being measured (von Davier, 2005a). Applying that to an educational assessment, Leighton and Gierl (2007) defined cognitive models in educational assessment as the “simplified description of human problem solving on standardized educational tasks, which helps to characterize the knowledge and skills students at different levels of learning have acquired and to facilitate the explanation and prediction of students’ performance” (p. 6). This is accomplished as CDMs are a special class of latent models wherein a student’s ability estimate is used to model the probability of correctly answering an item (Henson & Douglas, 2005; Henson, Roussos, Douglas, & He, 2008). CDMs can identify specific skills a student has not mastered by linking item response performance with attributes, measured by the items on an assessment. For example, a diagnostic mathematics assessment built to identify areas of struggle for a student not only provides an educator with the number of items the student answered correctly, but what skills the student has not mastered that were linked to the probability of a correct response on the missed items.

By directly evaluating observed responses and identifying current student understanding, the estimation of CDMs is well suited for formative and summative educational assessments and particularly the questions surrounding how best to identify student success and confusion. The latent variables, again, noted as skills or attributes, within CDMs are predominantly categorical and are typically dichotomous, allowing a student to be classified as master or non-master by evaluating the relationship between observed data and the set of latent variables (Templin & Henson, 2006; Templin et al.). The value of CDMs in education is obvious and clear; provide additional information for educators to ascertain students' current level of understanding and deficits to assist and guide remediation when necessary next step instructional activities (Tatsuoka). By harnessing the estimation of these models, the results are information that can then be used to create student education plans that focus efforts on specific skills a student is struggling to succeed with, providing feedback on where the student is currently and what work can progress understanding of the content.

When an attribute hierarchy is theorized to be present in the learning processes, it is assumed the student must progress through specified steps to successfully master the content. Cognitive diagnosis models have been designed to serve the purpose of identifying precise information about a student's level of understanding and the processes students pass through in completing assessment items under theorized learning hierarchies; however, should a hierarchy be present, conventional CDMs over fit the data by estimating all possible patterns of attribute mastery (2^k); meaning, if a hierarchy is present the model specifies item parameters that are redundant (Templin & Bradshaw, 2013). For example, an assessment measuring three attributes would result in the estimation of 8 attribute profiles ($2^k = 2^3= 8$). Should two of those attributes be dependent on the mastery of the other attributes being measured, a hierarchy exists, and there are now item parameters estimated for profiles that are not possible since the hierarchy stipulates the student must master the attributes in order of the hierarchy.

It is Templin and Bradshaw's (2013) work that addresses this over-specification by extending the Log-linear Cognitive Diagnostic Model (LCDM) to address cases where attribute hierarchies are present, therein establishing a link between LCDM and the functionality of the Attribute Hierarchy Model. Using a language acquisition assessment, their work presented the Hierarchical Diagnostic Classification Model (HDCM) as a credible model for statistically identifying attribute hierarchies that is absent in other latent class diagnostic models. The HDCM constrains the LCDM to a theorized hierarchy by setting some of the parameters within the model to zero. Referring to the previously mentioned three attribute assessment that would result in 2^3 = 8 attribute profiles using LCDM, there should be two attributes that are dependent on mastery of the other attribute (e.g., attribute 2 and 3 are dependent on attribute 1), the LCDM over specifies the data. Under the LCDM, item parameters for attribute 2 and 3 have been estimated although, based on the hierarchy, the attributes are dependent on attribute 1 so the profiles for attribute 2 and 3 that do not also include attribute 1 are not possible. In scenarios as this, the HDCM can be used to statistically test for the presence of a potential attribute hierarchy prior to subsequent analysis with models such as AHM (Templin & Bradshaw). By analyzing data using HDCM prior to AHM, the researcher can establish the presence of the theorized hierarchy and formulate the support for the critical assumption of AHM, which is that the hierarchy is in fact true.

This study was designed to evaluate the efficiency of the HDCM in identifying, or in this case, verifying a learning hierarchy. The current research dictates that there is need for statistical models capable of testing for the presence of learning hierarchies that can then be used to accurately and appropriately identify attribute hierarchies for use with models such as the Rule Space model or the AHM. To evaluate this efficiency and follow the work of Templin and Bradshaw (2013) in language acquisition assessment, this study evaluated the effectiveness of the HDCM in identifying the presence of a theorized hierarchy in the learning of slope in mathematics, a scenario intended to represent a common K12 education formative assessment practice. By designing an analysis that evaluated the data set first using the LCDM to ascertain initial fit and parameter estimation, the LCDM was constrained to the structure of the Foundational Concepts of Slope Assessment Hierarchy (FCSAH) to investigate the changes in model fit to determine if the FCSAH was present in the data.

The estimates for student classification were indeed similar for the LCDM and HDCM, which is not surprising; given the HDCM is simply a constrained version of the LCDM designed as a nested structure to represent the FCSAH. In both models, more than half of the students were classified as not having mastered any of the five attributes measured by the FCSA. The classification representing the mastery of all the attributes measured on the FCSA was the next largest. The interesting note is that the next largest mastery profile represented the students that had been classified as mastering attributes 2, 3, 4, and 5. This is in direct contrast to the structure of FCSAH, as the hierarchy indicates that attribute 1 must be learned before a student can progress to attribute 2, which is followed by any combination of attributes 3-5. These findings highlighted the necessity and value of the HDCM analysis to statistically evaluate the presence of a learning hierarchy underlying the structure of the FCSA. Considering students are classified as having either mastered or not mastered a skill (or a combination of skills) and that information is typically used to create remediation plans, future classroom lessons and build individualized student instruction plans, it is critical that a hierarchy be identified prior to model analysis that provides student classification. To establish the presence of a hierarchy, the relevant comparison pertains to model fit between the LCDM and the HDCM. As the data was gathered from an assessment designed using AHM and a specified attribute hierarchy was assumed, several profiles within the LCDM were constrained to zero in the HDCM model to align with the structure of the FCSAH. By removing the redundant profiles of the LCDM, it was expected that the model fit would improve as the model now represented the FCSAH hierarchy. However, the overall model fit suffered under the HDCM indicating a model constrained to the hierarchy erodes the model fit and the LCDM should be considered the superior fitting model. This erosion of fit indicates that there is reason to question the presence of the FCSAH attribute hierarchy.

Looking at the proportion of students classified to each mastery profile in the LCDM in comparison the previous AHM analysis highlights the differences in estimation that occur in unstable models. The LCDM classified most of the students as either masters of all or none of the attributes being measured on the FCSA. The previous AHM analysis of the FCSA (Broaddus, 2011) resulted in most of the students within the data set to be classified as having mastered many, if not all, of the attributes of the FCSAH. This is a stark difference that would benefit continued evaluation, not only with the current assessment data but across assessment research. It is this difference in classification values that is alarming when considered in real-world application. For example, assuming the output from the LCDM is correct, more than half of the students were classified has having mastered none of the attributes measured by the FCSA, whereas, the AHM analysis indicated that approximately 1% of the students were classified as having mastered none of the attributes. In a classroom scenario, an educator receiving the LCDM classification results would create very different learning and remediation plans than would a teacher receiving the AHM classification results. Understanding that substantial research has shown the most valuable feedback about and to a student should accurately identify where the student is in understanding the content and what are the next steps that would assist the student in reaching the end goal, there is value in statistically identifying the hierarchy prior to applying hierarchy models for student classification.

This study set out to evaluate the efficiency of the HDCM in statistically testing for an attribute hierarchy. Using results from a formative assessment designed using AHM based on a theorized hierarchy, this study examined the underlying structure of the attributes measured by the formative assessment and the classification differences between an AHM analysis and the current HDCM analysis. Because nonhierarchical diagnostic models over-fit the data by specifying item parameters and allowing for mastery profiles that cannot be present based on an identified hierarchy structure, it was theorized that the AHM should fit the data better than the HDCM; however, the LCDM fit the data the best. Understanding that the LCDM specifies all item parameters and allows for every possible attribute combination, if an attribute hierarchy is present, when the model is constrained to align with the structure of the hierarchy, fit should improve. This indicates that there is reason to question the presence or structure of the outlined attribute hierarchy. Considering the initial goal of this study to evaluate the efficiency of the HCM to statistically test for a hierarchy, it can be concluded the model does in fact provide a means for researchers to investigate the presence of a theorized hierarchy. Additionally, comparing the classification proportions of the original AHM analysis and this study, the value of statistically confirming the presence and structure of a hierarchy can be seen in the classification differences noted between the models.

Limitations

The assessment used for this study was measuring a narrow window of student learning; the prerequisites for working with slope. There is reason to consider the implications of such precision; though this a probable benefit when considering the application of analysis within formative assessment where focus is generally centered around a very limited number of skills, one must consider the effectiveness in broader assessment scenarios. Additionally, this study utilized mathematics assessment data due to availability, however, given the results of this study there is a considerable need to evaluate HDCM within all areas of student education including reading, writing, science and language. The sample size, though relatively large, would be another considered limitation as Templin and Bradshaw's (2013) initial work utilized a sample of almost 3,000 examinees.

Recommendations for Future Research

Understanding the utility of the HDCM, future research needs to evaluate the effectiveness of the model with large scale assessment data as well as interim and formative assessments. Being able to accurately test for a learning hierarchy provides a wealth of information for educators to then confidently apply response pattern models such as the Attribute Hierarchy model. If one were to consider the current state of the K12 industry, there is constant and consistent demand for tools and resources that can provide insight into current student understanding. By extending HDCM research across assessment applications, there is an opportunity to further understand its utility in today’s educational assessment research, particularly K12 formative systems that blend both assessment and instruction (Poggio & Meyen, 2009). Additionally, as the mandate for more diagnostic information in educational assessments can be expected to continue, there is a benefit to understanding more of the utility of the HDCM, particularly with more studies using assessments designed for student classification.

References

ARG. (2002). Assessment for Learning: 10 Principles., from www.assessment-reform-group.org.uk

Black, P. (2007). Full marks for feedback. Make the Grade: Journal of the Institute of Educational Assessors, 2(1), 18-21.

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment: Granada Learning.

Black, P., & Wiliam, D. (2009). Developing the Theory of Formative Assessment. Educational Assessment, Evaluation and Accountability, 21(1), 5-31.

Bloom, B. S., Hasting, H., & Madaus, G. F. (1983). Handbook of Formative and Summative Evaluation. New York, NY: McGraw-Hill.

Broaddus, A. (2011). AN INVESTIGATION INTO FOUNDATIONAL CONCEPTS RELATED TO SLOPE: AN APPLICATION OF THE ATTRIBUTE HIERARCHY METHOD. PhD Dissertation, University of Kansas, Lawrence, Kansas.

Close, C. N. (2012). *An Exploratory Technique for Finding the Q-matrix for the DINA Model in Cognitive Diagnostic Assessment: Combining Theory with Data.* UNIVERSITY OF MINNESOTA.

Cui, Y., & Leighton, J. P. (2009). The hierarchy consistency index: Evaluating person fit for cognitive diagnostic assessment. *Journal of Educational Measurement, 46*(4), 429-449.

de la Torre, J. (2008). An Empirically Based Method of Q-Matrix Validation for the DINA Model: Development and Applications. *Journal of Educational Measurement, 45*(4), 343-362.

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*(3), 333-353.

de la Torre, J., Hong, Y., & Deng, W. (2010). Factors affecting the item parameter estimation and classification accuracy of the DINA model. *Journal of Educational Measurement, 47*(2), 227-249.

de la Torre, J., & Karelitz, T. M. (2009). Impact of diagnosticity on the adequacy of models for cognitive diagnosis under a linear attribute structure: A simulation study. *Journal of Educational Measurement, 46*(4), 450-469.

DiBello, L. V., Roussos, L. A., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. *Handbook of Statistics Psychometrics, 26,* 979-1030.

DiBello, L. V., & Stout, W. (2007). Guest Editors' Introduction and Overview: IRT-Based Cognitive Diagnostic Models and Related Methods. *Journal of Educational Measurement, 44*(4), 285-291.

Embretson, S. E. (1983). Construct Validity: Construct Representation Versus Nomothetic Span. Psychological Bulletin, 93(1), 179-197.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists* (Vol. 4): Psychology Press.

Gierl, Leighton, J., & Hunka, S. M. (2007). Using the attribute hierarchy method to make diagnostic inferences about respondents' cognitive skills. In J. P. Leighton & M. J. Geirl (Eds.), *Cognitive diagnostic assessment for education: theory and application* (pp. 242-274). Cambridge: Cambridge University Press.

Gierl, M. J. (2007). Making Diagnostic Inferences About Cognitive Attributes Using the Rule-Space Model and Attribute Hierarchy Method. *Journal of Educational Measurement, 44*(4), 325-340.

Gu, Z. (2011). *Maximizing the Potential of Multiple-Choice Items for Cognitive Diagnostic Assessment.* University of Toronto.

Haertel, E. H. (2005). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, *26*(4), 301-321.

Hartz, S. M. C. (2002). *A Bayesian Framework for the Unified Model for Assessing Cognitive Abilites: Blending Theory with Practicality.* University of Illinois at Urbana-Champaign.

Henson, R., & Douglas, J. (2005). Test construction for cognitive diagnosis. *Applied Psychological Measurement*, *29*(4), 262-277.

Henson, R., Roussos, L., Douglas, J., & He, X. (2008). Cognitive diagnostic attribute-level discrimination indices. *Applied Psychological Measurement*, *32*(4), 275-288.

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

Henson, R., Templin, J., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74*(2), 191-210.

Huebner, A., Wang, B., & Lee, S. (2009). *Practical issues concerning the application of the DINA model to CAT data.* Paper presented at the Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing. Retrieved [date] from www. psych. umn. edu/psylabs/CATCentral.

Huebner, A. (2010). An Overview of Recent Developments in Cognitive Diagnostic Computer Adaptive Assessments. Practical Assessment, Research & Evaluation, 15(3), n3.

Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for Fusion Model application to LanguEdge assessment. *Language Testing, 26*(1), 031-073.

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

Kim, H. S. J. (2011). Diagnosing examinees' attributes-mastery using the Bayesian inference for binomial proportion: a new method for cognitive diagnostic assessment.

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

Leighton, J. P., & Gierl, M. J. (2007). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educational Measurement: Issues and Practice, 26*(2), 3-16.

Liu, J., Xu, G., & Ying, Z. (2010). Theory of Self-learning Q-Matrix. *arXiv preprint arXiv:1010.6120*.

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

Nichols, P. D., Chipman, S. F., & Brennan, R. L. (2012). *Cognitively diagnostic assessment*: Routledge.

Pellegrino, J. W., Baxter, G. P., & Glaser, R. (1999). Addressing the" two disciplines" problem: Linking theories of cognition and learning with assessment and instructional practice. *Review of research in education, 24*, 307-353.

Poggio, J., & Meyen, E. (2009). Blending Assessments with Instruction Program: A formative system. Paper presented at the American Educational Research Association, San Diego, CA.

Ramaprasad, A. (1983). On the definition of feedback. Behavioral Science, 28(1), 4-13.

Roussos, L. A., DiBello, L. V., Stout, W., Hartz, S. M., Henson, R. A., & Templin, J. (2008). The fusion model skills diagnosis system. *Cognitive diagnostic assessment for education: Theory and applications*, 275-318.

Rupp, A., & Mislevy, R. J. (2007). Cognitive foundations of structured item response theory models. In J. Leighton & M. Gierl (Eds.), *Cognitive diagnostic assessment in education: Theory and practice* (pp. 205-241). Cambridge: Cambridge University Press.

Rupp, A., & Templin, J. (2008a). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement, 68*(1), 78-96.

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

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

Scriven, M. S. (1967). The methodology of evaluation (Perspectives of Curriculum Evaluation, and AERA monograph Series on Curriculum Evaluation, No. 1). Chicago: Rand NcNally.

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 and Behavioral Statistics, 10*(1), 55-73.

Tatsuoka, K. K. (1993). Item construction and psychometric models appropriate for constructed responses. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 107-133). Hillsdale, NJ: Erlbaum.

Tatsuoka, K. K. (2009). *Cognitive assessment: An introduction to the rule space method*: CRC Press.

Templin, J. (2006). *CDM user's guide*. Unpublished manuscript.

Templin, J., & Bradshaw, L. (2013). Hierarchical Diagnostic Classification Models: A Family of Models for Estimating and Testing Attribute Hierarchies. *Psychometrika, 30*(2), 251-275.

Templin, J., & Hoffman, L. (2013). Obtaining Diagnostic Classification Model Estimates Using Mplus. Educational Measurement: Issues and Practice, 32(2), 37-50.

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

Templin, J., Henson, R., Templin, S., & Roussos, L. (2008). Robustness of hierarchical modeling of skill association in cognitive diagnosis models. *Applied Psychological Measurement, 32*(7), 559-574.

Topol, B., Olson, J., Roeber, E., & Hennon, P. (2012). *Getting to higher-quality assessments: Evaluating costs, benefits, and investment strategies*. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

von Davier, M. (2005a). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61*(2), 287-307.

von Davier, M. (2005b). mdltm: Software for the general diagnostic model and for estimating mixtures of multidimensional discrete latent traits models [Computer software]. Princeton, NJ: ETS.

Wiliam, D., & Thompson, M. (2007). Integrating assessment with instruction: what will it take to make it work? In C. A. Dwyer (Ed.), The future of assessment: shaping teaching and learning (pp. 53-82). Mahwah, NJ: Lawrence Erlbaum Associates.

Xu, X., Chang, H., & Douglas, J. (2003). *A simulation study to compare CAT strategies for cognitive diagnosis.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Zhou, J., Gierl, M. J., & Cui, Y. (2009). *Attribute Reliability in Cognitive Diagnostic Assessment*. Paper presented at the National Council on Measurement innEducation, San Diego.